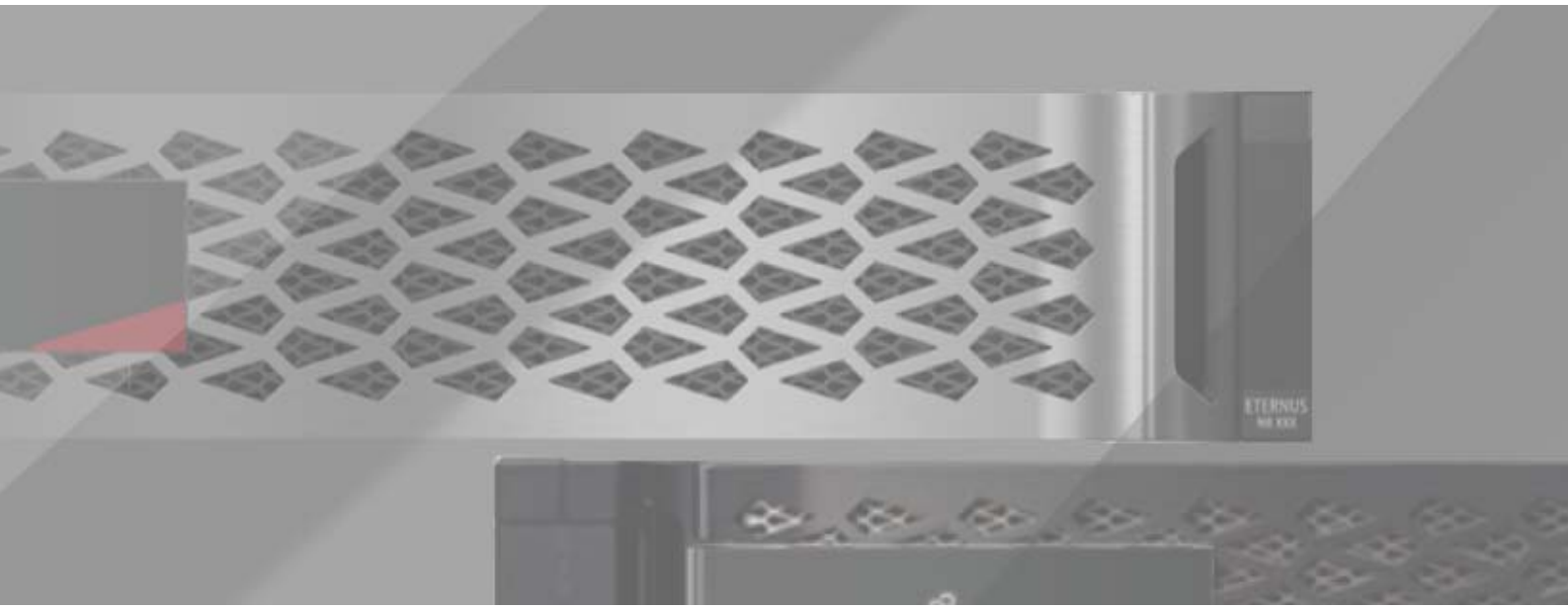
C140-0044-01Z3

Fujitsu Storage
ETERNUS AX series オールフラッシュアレイ ,
ETERNUS HX series ハイブリッドアレイ

NVMe-oF を使用した最新の SAN の実装と構成



FUJITSU

目次

図目次

表目次

はじめに

本書は、NVMe-oF トランスポート (NVMe/FC および NVMe/TCP) の実装および設定方法について説明しています。本書の記載には、NVMe プロトコルとトランスポートを使用して可用性とパフォーマンスに優れた最新の SAN ソリューションを構築するための、設計、実装、構成、管理のガイドラインとベストプラクティスが含まれます。

初版
2023 年 6 月

登録商標

本製品に関連する他社商標については、以下のサイトを参照してください。
https://www.fujitsu.com/jp/products/computing/storage/trademark/

本書では、本文中の ™、® などの記号は省略しています。

本書の読み方

対象読者

本書は、ETERNUS AX/HX の設定、運用管理を行うシステム管理者、または保守を行うフィールドエンジニアを対象としています。必要に応じてお読みください。

関連マニュアル

ETERNUS AX/HX に関連する最新の情報は、以下のサイトで公開されています。
https://www.fujitsu.com/jp/products/computing/storage/manual/

本書の表記について

■ 本文中の記号

本文中では、以下の記号を使用しています。

注 意	お使いになるときに注意していただきたいことを記述しています。必ずお読みください。
備 考	本文を補足する内容や、参考情報を記述しています。

第 1 章
NVMe とは

NVMe (NVM Express データストレージ標準) は、新しいストレージインフラストラクチャの構築や最新のインフラストラクチャへのアップグレードを行う企業向けのコアテクノロジーとして台頭しています。

NVMe は、ソリッドステートストレージデバイス向けに最適化されたプロトコルであると同時に、NVMEM のコンポーネントやシステム向けのオープンソースアーキテクチャ標準でもあります。

NVMe 標準は、現在および将来のメモリテクノロジー向けに、高帯域幅で低レイテンシのストレージアクセスを提供するように設計されています。NVMe は、SCSI コマンドセットを NVMe コマンドセットで置き換え、SCSI、Serial Attached SCSI (SAS) 、SATA などの従来の標準よりもはるかに高速広帯域幅のハードウェアプロトコルである、PCIe に依存します。

SCSI はおよそ 40 年前に導入され、当時普及していたストレージ技術 (8 インチフロッピーディスクやファイルキャビネットサイズ HDD) 向けに設計されました。また、当時利用可能だった非常に低速なシングルコア CPU とこれより少量の DRAM に合わせた設計でした。

一方、NVMe は、マルチコア CPU とギガバイトサイズのメモリで駆動する不揮発性フラッシュドライブで動作するように開発されました。1970 年代以降のコンピュータサイエンスの著しい進歩を利用しており、データをより効率的に解析および操作するための合理化されたコマンドセットを可能にしています。

NVMe は様々な用途に使われてきました。NVMe over Fabrics (NVMe-oF) の導入においては、メディアタイプの置き換えを検討しているかどうかを明確にすることが重要です。ほとんどの場合、あるメディアタイプを NVMe に接続されたドライブに交換することを意味します (図 1.1) 。この使用例では、ノート PC、デスクトップ、サーバー、さらにはストレージアレイなど、あらゆる既存のメディアタイプを NVMe に接続されたドライブで置き換えることができます。

図 1.1　NVMe の定義

1.1 NVMe 標準の歴史について

NVMe 標準バージョン 1.0 は、2011 年 3 月に NVM Express 社によって承認されました。以来、数回のアップデートが行われ、2017 年 11 月に承認された NVMe バージョン 1.3a が最新となっています。NVMe 社 (NVM Express 社) は、2015 年 11 月に補足的な仕様である NVMe Management Interface (NVMe-MI) を公開しました。NVMe-MI は、帯域内および帯域外管理の両方に重点を置いています。NVMe-oF 仕様は、2016 年 6 月に追加されました。NVMe-oF は、ネットワークまたはファブリック上の NVMe プロトコルを使用して定義されています。

NVMe 2.0 仕様ファミリーは、NVMe 仕様の現行バージョンです。NVMe-oF、NVMe-MI、NVMe-KV いった NVMe の拡張仕様に、承認済の技術提案 (TP) を組み合わせて、新たな NVMe 2.0 仕様ファミリーとしています。NVMe 標準は、最新のストレージデバイスのデバイスアクセスまでのソフトウェアスタックを対象としています。NVMe 2.0 の仕様と新しい技術提案については、NVM Express 社の「Everything You Need to Know About NVMe Specifications and New Technical Proposals」を参照してください。

備 考

技術提案は、Internet Engineering Task Force (IETF) の Request For Comments (RFC) の NVMe(NVM Express) 版です。技術提案は、技術提案がフォーカスする分野をカバーする NVM Express ワークグループにレビューに出されます。
最終的に技術提案は投票にかけられ、NVMe の仕様とプロセスの拡張機能として承認される可能性があります。

1.1.1 NVMe プロトコルの開発者と管理者について

NVMe は、Peripheral Component Interconnect (PCI) の標準化機構である Peripheral Component Interconnect Special Interest Group (PCI-SIG) から生まれた標準化機構によって開発されています。NVMe の標準化機構は、Non-Volatile Memory Express 社です。

1.1.2 NVMe 標準文書の入手先について

NVMe の仕様、ホワイトペーパー、プレゼンテーション、動画、その他の資料は、NVM Express 社の Web サイトからダウンロードできます。

NVMe/FC 規格は、International Committee for Information Technology Standards (INCITS) の Fibre Channel Industry Association (FCIA) による、T11 委員会の FC-NVMe 標準 (T11-2017-00145-v004「FC-NVMe」) の中で、詳細に定義されています。

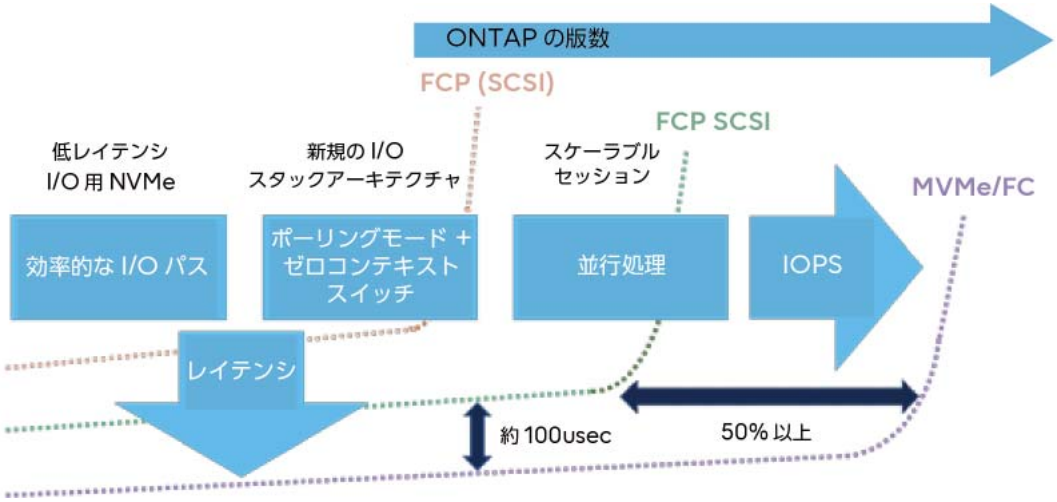
1.2　高速な NVMe-oF

NVMe は、データストレージのパフォーマンスにおける 4 つの重要な属性 (IOPS、スループット、レイテンシ、CPU 使用率) に対応することから、現代のデータセンターに不可欠な要素となりえます。

- **IOPS** は、デバイスが 1 秒あたりに実行できる読み取りまたは書き込み操作の数を表します。IOPS の値は、操作が読み取りであるか書き込みであるか、アクセスタイプがシーケンシャルかランダムかによっても異なり、保存または取得されるブロックサイズによっても変わります。ほとんどのデバイスでは、4kB または 8kB などの小さい I/O サイズのブロックを使用すると、IOPS 値が高くなります。しかし、実際のアプリケーションでは、32kB や 64kB など、より大きな I/O サイズのブロックが必要になることが多いため、関連する I/O 特性に基づいて評価することが重要です。
- **スループット**は、ストレージデバイスがデータを読み取りまたは書き込みできる速度の指標で、通常は GB/s で指定されます。通常、I/O サイズが大きいほど高くなりますが、I/O の方向とアクセスタイプ (ランダムまたはシーケンシャル) によっても異なります。したがって、この場合も、実際の運用環境を考慮して評価する必要があります。
- **レイテンシ**は、読み取りまたは書き込み操作の開始から完了までの時間です。ストレージのレイテンシは、転送されるデータのサイズや、アクセスタイプがシーケンシャルかランダムか、操作が読み取りか書き込みかによっても変わるほか、ネットワークの速度によっても異なります。ストレージのレイテンシ、特に読み取りのレイテンシを小さくすることは、応答性と魅力的なエクスペリエンスをユーザーに提供するために不可欠です。
- **CPU 使用率**は、レイテンシ内でまたは十分なスループットを維持して、(1 つまたは複数の) 既存ワークロードを達成するのに必要な I/O の生成に使用される、CPU サイクルの数を示す指標です。上記の他の指標とは異なり、NVMe の効率性により CPU 使用率が低下することは、FCP や iSCSI などの SCSI ベースのプロトコルから NVMe に移行するメリットとしてあまり知られていません。CPU 使用率を削減するメリットとしては、既存のストレージコントローラー上でより多くのワークロードを統合できることや、特定のワークロードをホストするために必要なサーバの数を削減できる可能性があります。ストレージ側とホスト側の両方の削減により、IT 投資が増加し、NVMe に移行するプロジェクトの投資収益率 (ROI) が急速に向上します。

図 1.2 および図 1.3 に、なぜ NVMe がそこまで効果的かつ高速であるかの理由を示します。

図 1.2　NVMe/FC が高速である理由



IOPS と帯域幅の向上は、主に NVMe の柔軟性と、高速転送テクノロジーを利用した NVMe のコマンドとデータの移行機能の結果です。これらの転送には以下のものがあります。

- **FCP**
  現在は 32Gbps と 64Gbps の速度で提供されています。将来的に 128Gbps がサポートされる予定です。
- **リモートダイレクトメモリアクセス (RDMA) プロトコル**
  - データセンターの高速 Ethernet：現在、25、40、50、100、および 200Gbps で利用可能です。
  - InfiniBand (IB)：現在、最大 100Gbps の速度で利用可能です。
- **PCI Express 3.0**
  1 秒あたり 8 ギガの転送 (GTps) 、つまり約 6.4Gbps をサポートします。
- **NVMe コマンドおよびペイロードを許容する TCP によるユビキタスな Ethernet ネットワークの使用**
  現在、10、25、40、50、100、さらに 200Gbps の高速 Ethernet ネットワークが使用可能です。

パフォーマンスの改善は、NVMe が可能にした大規模な並列化の結果です。この並列化により、プロトコルは複数のスレッドの同時処理のために複数のコアに処理を分散できます。

レイテンシの改善は、以下のような要因の組み合わせの結果です。

- 高い並列処理。I/O 送信キューと完了キューのペアは、ホスト CPU コアに配列されます。各ホスト / コントローラーペアには、NVMe キューの独立したセットがあります。
- NVMe コマンドセットの合理化
- ハードウェア割り込みを置き換えるポーリングモードドライバー
- ソフトウェアロックの排除
- コンテキストスイッチの削除

これらの要素が連携して、企業のビジネスクリティカルなアプリケーションの重要な指標であるスループットの向上とレイテンシの削減を実現しています。

図 1.3　NVMe プロトコルが効率的である理由

1.3　ストレージ接続アーキテクチャとしての NVMe

NVMe の現在の主な使用用途は、ドライブやドライブシェルフの接続です。多くのストレージベンダーやサプライヤーは、ストレージ接続のアーキテクチャおよび規格として NVMe を使用した製品を導入しています。技術的には、NVMe が I/O の実行に使用されるプロトコルであるのに対し、基本的な物理転送は PCIe であることがほとんどです。

このシナリオでは、NVMe は SCSI コマンドセットを NVMe コマンドセットで置き換え、ドライブをストレージコントローラーに接続するために SATA または SAS を PCIe で置き換えることがよくあります。NVMe は物理的な接続と転送に依存します。転送に PCIe を使用します。

NVMe 接続フラッシュでは、以下の理由により、帯域幅が増加し、レイテンシが減少します。

- より多く、深い深度のキューの提供 64k (65,535) キューを提供し、各キューの深度は 64k です。
- NVMe コマンドセットは合理化されているため、従来の SCSI コマンドセットよりも効率的です。

NVMe コマンドセットを使用して、SAS 12GB バックエンドおよび SCSI コマンドセットを PCIe 接続ドライブに変更すると、あらゆるバックエンドプロトコルにおいて、パフォーマンス (スループット) が向上し、レイテンシが減少します。この改善は、より効率的で、必要なプロセッサパワーが少なく、並列化が可能なドライブアクセスによるものです。理論的には、パフォーマンスの向上により、スループットが約 10 ～ 15% 向上し、レイテンシが 10 ～ 25% 削減されます。ワークロードプロトコル、コントローラー上で実行される他のワークロード、および I/O を実行するホストの相対的なビジー状態の違いによって、これらの値は明らかに大きく変化します。

1.4　エンドツーエンドの NVMe：ETERNUS AX4100

ETERNUS AX4100 は、NVMe 接続のソリッドステートドライブ (SSD) を採用したストレージアレイです。

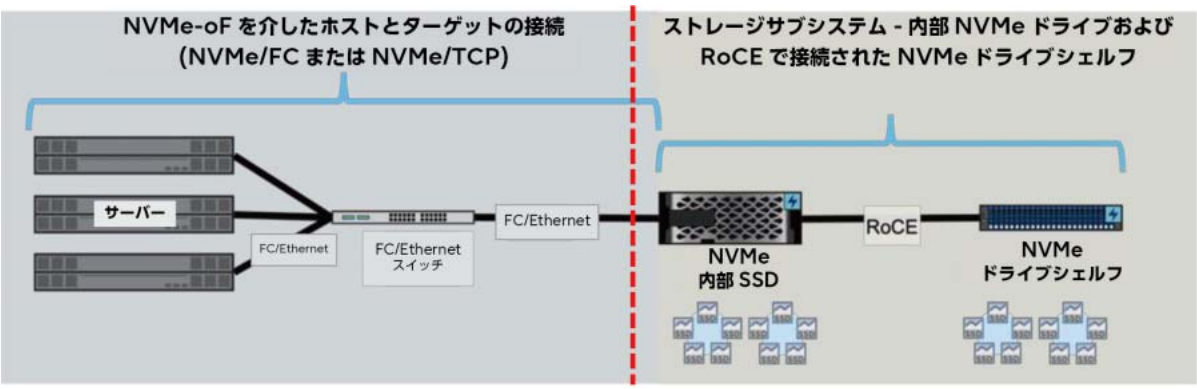
図 1.4　ETERNUS AX4100

ここでは、いくつかの注目すべきポイントを示します。

- 業界初の、32Gbps FC を介したエンドツーエンドの NVMe/FC ホスト間フラッシュアレイ。AX4100 の FC ホストバスアダプタ (HBA) は、16 または 8Gbps (N-2) まで自動ネゴシエートできます。
- 業界初の、NAS および iSCSI プロトコルのフロントエンド接続用の 100GbE 接続。
- 15.3TB NVMe SSD を搭載した 4U シャーシで 2.5PiB の有効容量 (サポート予定)。
- 100GbE MetroCluster IP (MCC IP) によるピークパフォーマンス。

高性能かつ低レイテンシの内蔵 NVMe SSD による、あらゆるアプリケーションの高速化。さらに、AX4100 を NVMe/FC と組み合わせると、エンドツーエンドの NVMe 接続が実現します (図 1.5) 。これにより、組織はより低いレイテンシでより高いパフォーマンスを実現できます。

図 1.5　エンドツーエンド NVMe



一部のお客様は、100GbE をサポートするストレージソリューションを期待しています。また、NAS および iSCSI プロトコル用の 100GbE と、FCP および NVMe/FC 用の 32Gbps FC の両方を有効にして、最終的に優れたアプリケーション帯域幅を提供するシステムを導入することも考えられます。

AX4100 は、SAN のみ、NAS のみ、またはその両方の環境に最適です。

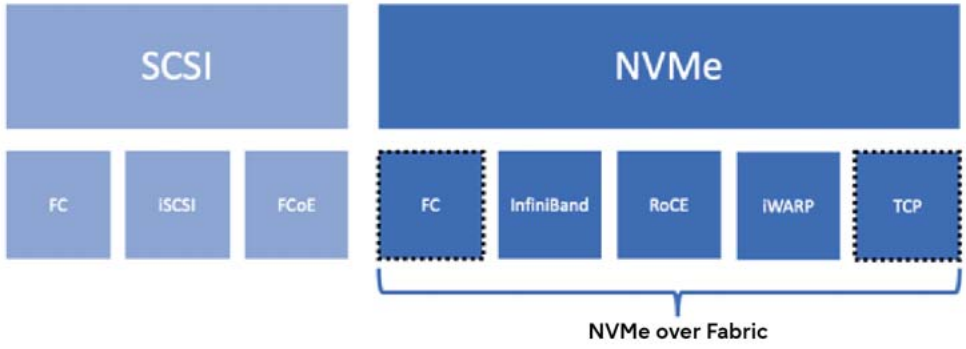
第 2 章
NVMe-oF

NVMe は単なるストレージの仕様ではありません。NVMe-oF プロトコル拡張には、サーバーからネットワーク、ストレージシステムに至るデータパスの全体が含まれます。HDD を SSD に置き換えることによって HDD のボトルネックが解消された後、別のボトルネックが発生しました。ローカルと SAN の両方でデータにアクセスするために使用されるストレージプロトコルです。

NVMe over Fabrics 委員会は、ローカルストレージのアクセスプロトコルをアップグレードする、NVMe を導入しました。委員会は、NVMe-oF の発売に合わせて、NVMe プロトコルおよびコマンドセットをさまざまなネットワークおよびファブリックプロトコルを使用して長距離で使用するための仕様とアーキテクチャを追加しました。その結果、パフォーマンスが大幅に向上し、FCP や iSCSI から NVMe/FC などの NVMe-oF 仕様に移行したワークロードのレイテンシが削減されました。

エンドツーエンドの NVMe を実装するには、NVMe に接続されたソリッドステートメディアだけでなく、ストレージコントローラーからホストサーバーへの NVMe 転送も必要です。オリジナルの NVMe 仕様は、コマンドセットと接続のアーキテクチャを設計しましたが、転送方式と送信方式は PCIe (またはそのほかの物理転送仕様) に依存します。この仕様では、主に不揮発性フラッシュストレージ技術をローカルサーバに接続することに焦点を当てていました。NVMe は SCSI コマンドを置き換え、処理キューの数と各処理キューのキュー深度を増加させます。NVMe はコンテキストスイッチを減らし、ロックレスです。これらの機能強化により、PCI バスを使用する NVMe 接続ドライブなど、NVM に接続されたドライブのアクセス時間とレスポンスタイムが大幅に向上します。図 2.1 は、SCSI と NVMe-oF のプロトコルと転送方式の組み合わせをいくつか比較したものです。

図 2.1　SCSI と NVMe-oF の転送方式の比較



2.1 NVMe およびデータファブリック

NVMe は、ローカルの不揮発性ストレージをコンピュータまたはサーバに接続するためのアクセスプロトコルとアーキテクチャを定義します。NVMe-oF は、スケーリングとレンジの改善を追加することで、元の NVMe 仕様を強化します。NVMe-oF は、FC、Ethernet、IB といった様々なネットワークストレージの転送方式を介した NVMe の転送仕様を定義し作成することで、SAN 市場に NVMe を導入させる効果的な NVMe の拡張機能です。

最終的に、NVMe-oF は NVMe を新しいブロックストレージプロトコルタイプとして追加します。NVMe/FC などの特定の NVMe-oF トランスポートを開発する場合は、ベンダーが従うべき転送プロトコルとアーキテクチャの仕様をまとめて指定します。

NVMe-oF は、NVMe が FC や Ethernet などの既存の転送技術を使用して、長距離で NVMe プロトコルを転送し、スイッチやルータなどのネットワークテクノロジーを使用できるようにする方法を定義します。これらの転送プロトコルをサポートすることで、NVMe-oF は大規模ストレージアレイのパフォーマンスを大幅に向上させると同時に、以下のように他のプロトコルと置き換えることでストレージプロトコルの並列化を強化します。

- **FCP**
  FC フレームにカプセル化された SCSI コマンドです。
- **iSCSI**
  IP/Ethernet フレームにカプセル化された SCSI コマンドです。
- **FCoE**
  FC フレームにカプセル化された SCSI コマンドを、さらに Ethernet フレームにカプセル化しています。
- ONTAP 9.10.1 は、Ethernet ベースの NVMe-oF 転送方式として NVMe/TCP を追加しました。
- 図 1.5 は、NVMe-FC と新しい AX4100 を使用して、エンドツーエンドの NVMe フラッシュを提供する ONTAP を示しています。

NVMe-oF は、主に NVMe プロトコルをデータネットワークおよびファブリックに拡張することを目的としています。これは、コンピュータをブロックベースのストレージに接続するために使用されるアクセスアーキテクチャとプロトコルを定義します。以下のような現在のブロックプロトコルの更新と考えるのが最も簡単です。

- **FCP**
  FCP は、FC フレーム内に SCSI コマンド記述ブロック (CDB) をカプセル化します。FC は転送方式を定義しますが、FCP は特に FC プロトコルを使用して SCSI (CDB) をカプセル化することを意味します。現在、FCP は最も一般的な SAN プロトコルです。FCP ファブリック (ネットワーク) の速度は、1 ～ 32 Gbps です。8 Gbps、16 Gbps、および 32 Gbps が最も頻繁に使用される速度です。

- **iSCSI**
  iSCSI は、Internet Engineering Task Force (IETF) によって、RFC 3270 Internet Small Computer Systems Interface の中で最初に定義されました。RFC 3270 は、後継となる RFC 7143 Internet Small Computer System Interface (iSCSI) Protocol (Consolidated) に置き換わっています。RFC 7143 は、2004 年発表の RFC 3270 の仕様を元にアップデートおよび改版したものです。

NVMe-oF は、現存するさまざまな転送方式で NVMe を転送するために使用できる仕様、アーキテクチャ標準、およびモデルを提供します。NVMe-oF の転送方式には、以下が含まれます。

- **NVMe/FC**
  転送方式に FC を使用する NVMe です。詳細については、「2.3 NVMe over Fibre Channel」(P.24) を参照してください。

- **NVMe が TCP データグラムにカプセル化された NVMe/TCP**
  NVMe/TCP は、最も一般的な NVMe over Ethernet のバリエーションとなる可能性があり、最終的には iSCSI の論理的な代替となる可能性があります。iSCSI と同様に、NVMe/TCP は標準の NIC と Ethernet スイッチを使用します。このため、RDMA NIC (RNIC) などの特別に構築されたブロックや、RDMA over Converged Ethernet(RoCE) をサポートするためのデータセンターブリッジング (DCB) スイッチが不要で、これから NVMe-oF on Ethernet を導入する環境に適したオプションになります。

- **RDMA を用いた NVMe 転送方式**
  RDMA をサポートする転送方式はいくつか存在しています。
  - NVMe over InfiniBand (NVMe/IB)
    このソリューションでは、現在 100 Gbps をサポートできる IB を超高速転送方式として使用します。IB は非常に高速ですが、高価であり、距離と規模の両方に制限があります。NVMe/IB ターゲットを使用して NVMe-oF を提供する最初のエンタープライズクラスのストレージアレイは、ETERNUS AB5100 です。ETERNUS AB5100 は、2U プラットフォームで 1M IOPS と 21GBps を 100ms 未満で実現します。

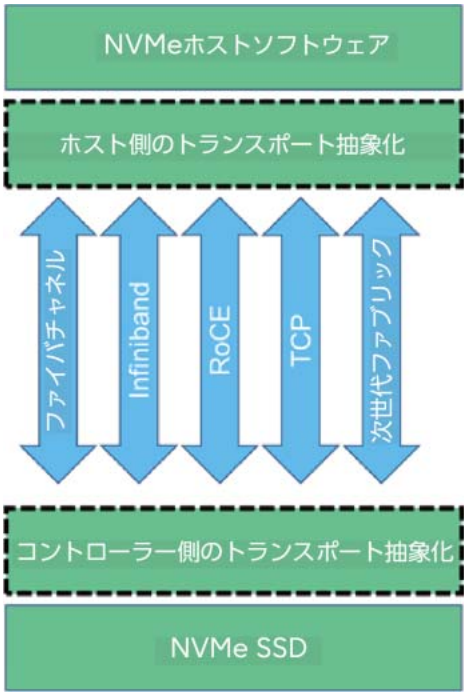
  - **RDMA over Converged Ethernet (RoCE)**
    - Internet Wide-Area RDMA Protocol (iWARP) は、TCP または Secure TCP (STCP) を使用して転送される Direct Data Placement Protocol (DDP) を使用して RDMA を転送します。DDP はデータをストリームで送信し、TCP プロトコルのデータユニットにデータを収めるためのセグメント化をしません。
    - RoCE は TCP を必要としないため、レイテンシが低くなります。RoCE には、DCB およびプライオリティフロー制御 (PFC) をサポートする Ethernet スイッチが必要です。DCB スイッチは、標準の Ethernet スイッチとは異なり、高価になる傾向があります。ホストとストレージコントローラーには、RNIC がインストールされている必要があります。これらの要件により、クラウドへの RoCE の採用が制限される場合があります。RoCE には 2 つのバリエーションがあります。
      - RoCE v1 (元の RoCE 仕様) は、同一サブネット内のイニシエータとターゲット間の通信を可能にするデータリンク層プロトコルを定義します。RoCE は、サブネット間でルーティングできないリンク層プロトコルです。

- RoCE v2 は、IPv4 または IPv6 上で User Datagram Protocol (UDP) を使用するインターネット層のプロトコルです。これは、サブネット間でルーティングできるレイヤ 3 のインターネット層プロトコルです。UDP では順序どおりの配信が強制されませんが、RoCE v2 仕様では順序どおりでないパケット配信が許可されていないため、DCB ネットワークは送信された順序でパケットを配信する必要があります。また、RoCE v2 は、Explicit Congestion Notification (ECN) ビットを用いてフレームと Congestion Notification パケット (CNP) をマーキングすることで受領を認証する、フローコントロール機構を定義します。

- **iWarp 上の RDMA**
  RDMA の使用をサポートするための拡張機能です。

図 2.2 は、NVMe-oF スタックと、転送に使用できるネットワークおよびファブリックの一部を示しています。

図 2.2　複数のネットワーク転送に使用できる NVMe



NVMe は、NVMe-oF も参照できます。NVMe-oF は、さまざまな一般的なネットワーキングおよびファブリックプロトコルの内部に NVMe コマンドセットおよびデータペイロードをカプセル化する方法を指定する標準拡張機能です。図 2.3 は、NVMe-oF スタックおよびこれらのプロトコル / 転送方式のいくつかを示します。また、NVMe-oF を使用して、NVMe-oF とネットワーク接続できるオブジェクトの直径および数を拡張する方法を示しています。

図 2.3　NVMe over Fabrics (NVMe-oF)

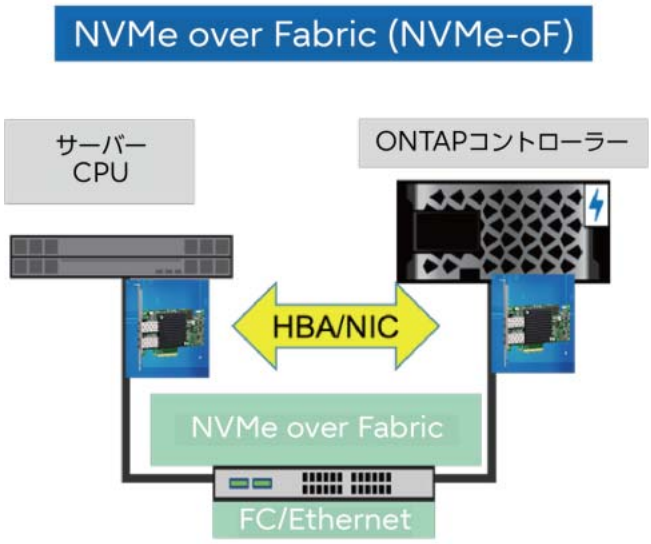


図 2.4 は、ネイティブ FC フレームと、FC フレーム内にカプセル化された NVMe の違いを示しています。このとおり、NVMe/FC は SCSI-3 コマンド記述子ブロック (CDB) を NVMe プロトコルコマンドで置き換えます。この単純な交換により、NVMe プロトコルを FC ファブリック内で転送できます。これにより、ネットワークに接続できるオブジェクトの直径と数が増加し、非常に信頼性の高い転送が可能になります。

図 2.4　FCP と NVMe/FC フレームの比較

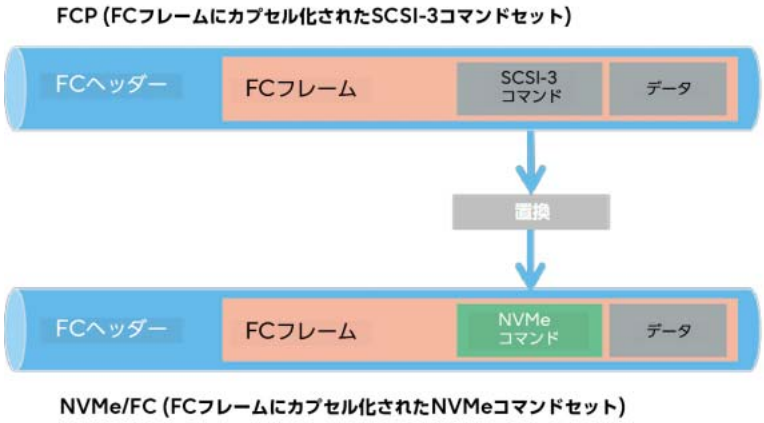


図 2.5 は、カプセル化され、TCP データグラムの内部で転送される NVMe コマンドセットである NVMe/TCP を示しています。

図 2.5　iSCSI と NVMe/TCP データグラムの比較

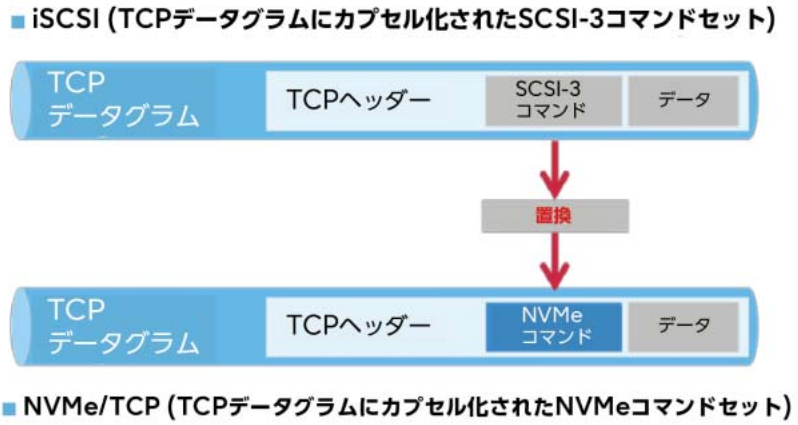


図 2.6 は、NVMe/TCP の別の図です。NVMe プロトコルとペイロードの両方が TCP データグラムのデータ / ペイロード部にカプセル化されていることを示します。このとおり、NVMe-oF は、さまざまな既存のネットワークプロトコルまたはファブリックプロトコルの内部に NVMe コマンドとデータペイロードをカプセル化して、NVMe を拡張し、それを遠隔地に転送します。

図 2.6　TCP データグラム内の NVMe

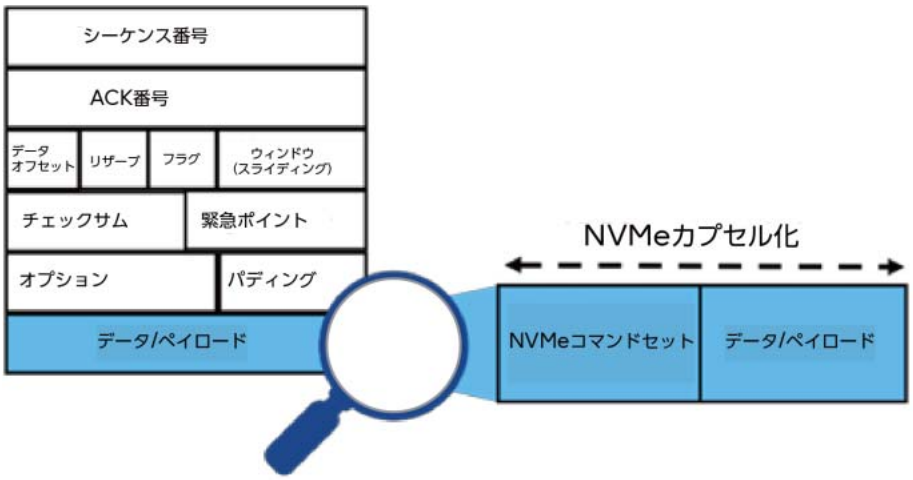
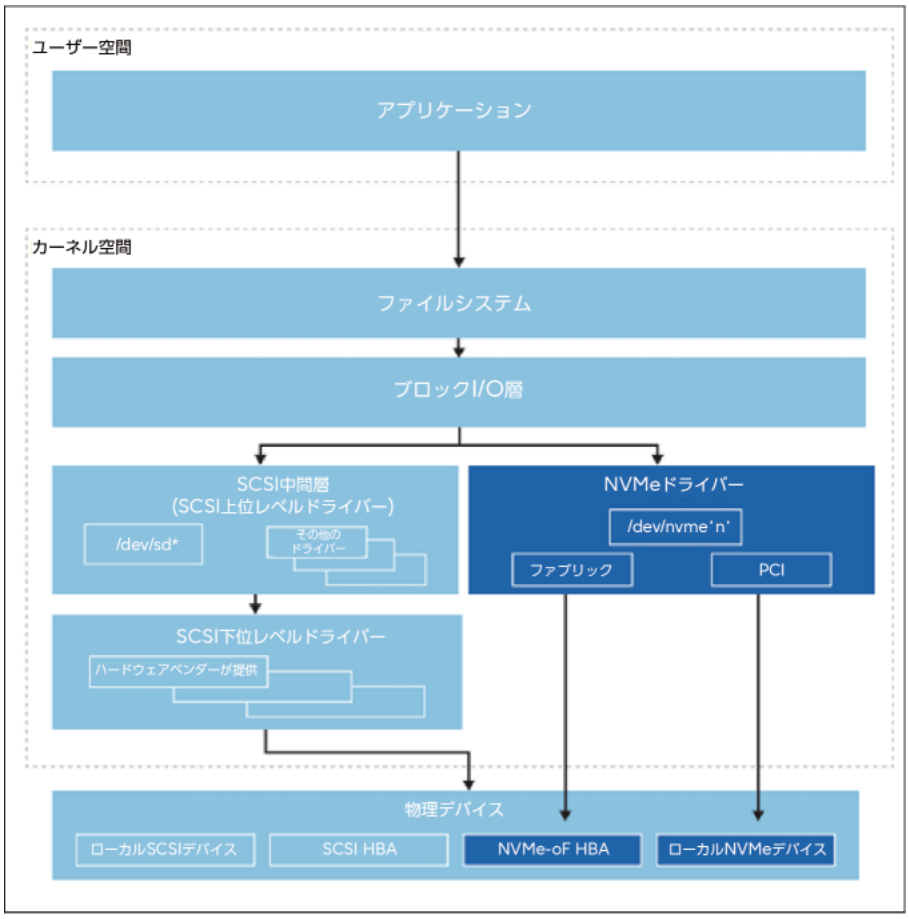


図 2.7 は、SCSI スタックと NVMe プロトコルスタックの比較です。NVMe スタックが SCSI スタックと比べてどれだけ短いかに留意してください。

図 2.7　SCSI スタックと NVMe スタックアーキテクチャの比較



2.1.1　オープンソースフレームワークの活用

いくつかのオープンソースのアーキテクチャ仕様を利用して、ONTAP NVMe-oF ターゲットを迅速に開発し、他の NVMe ハードウェアおよびソフトウェアとのシームレスな相互運用性をサポートしました。また、これらの仕様を採用し準拠させる一方で、NVM Express 社や INCITS などのいくつかの標準化機構にも積極的に参加しています。

- データプレーン開発キット (DPDK)
- ストレージパフォーマンス開発キット (SPDK)

2.2　NVMe と高可用性

NVM Express 委員会の代表である Fred Knight 氏が、機能的な高可用性エラーレポートとフェイルオーバープロトコルを定義する TP 4004 と TP 4028 に関する技術提案を提出しました。新しいプロトコルである Asymmetric Namespace Access (ANA) が 2018 年 3 月に承認されました。

2.2.1　ANA

ALUA と同様に、ANA では、ホスト側のマルチパス実装が各オペレーティングシステムスタックで使用されるストレージ HA マルチパスソフトウェアと連携するために必要なすべてのパスおよびパス状態情報を提供できるように、イニシエータ側とターゲット側の両方に実装が必要です。ANA が機能するには、ターゲットとイニシエータの両方が ANA を実装し、サポートしている必要があります。

NVMe/FC は、ANA プロトコルを使用してパスとターゲットの両方のフェイルオーバーに必要なマルチパスとパス管理を提供します。ANA プロトコルは、NVMe サブシステムがパスおよびサブシステムエラーをホストに返す方法を定義して、ホストがパスの管理およびパス間のフェイルオーバーの管理をできるようにします。NVMe/FC での ANA の役割は、ALUA が FCP プロトコルと iSCSI プロトコルの両方で果たす役割と同じです。MPIO や Device Mapper Multipathing (DM-Multipath) などのホストオペレーティングシステムのパス管理機能を備えた ANA は、NVMe/FC 向けのパス管理機能とフェイルオーバー機能を提供します。図 2.8 および図 2.9 は、承認のために NVM Express 社に提出された技術提案の表紙です。図 2.10 は、INCITS T11 が FC-NVMe 仕様を説明する T11-2017-00145-v004 の表紙です。INCITS T11 委員会は、Intelligent Peripheral Interface (IPI) 、High-Performance Parallel Interface (HIPPI) 、および FC の分野における規格開発を担当しています。

図 2.8　TP 4004：ANA ベースの提案書 (3/18 承認)

NVM Express Technical Proposal for New Feature

Technical Proposal ID	**TP 4004**
Change Date	**02/26/2018**
Builds on Specification	**NVM Express 1.3 or later; does not apply to versions earlier than 1.3.**

Technical Proposal Author(s)

Name	Company
Fred Knight	NetApp
David Black	Dell EMC
Curtis Ballard	HPE
Christoph Hellwig	WDC

図 2.9　TP 4028：ANA パスおよび転送 (1/18 承認)

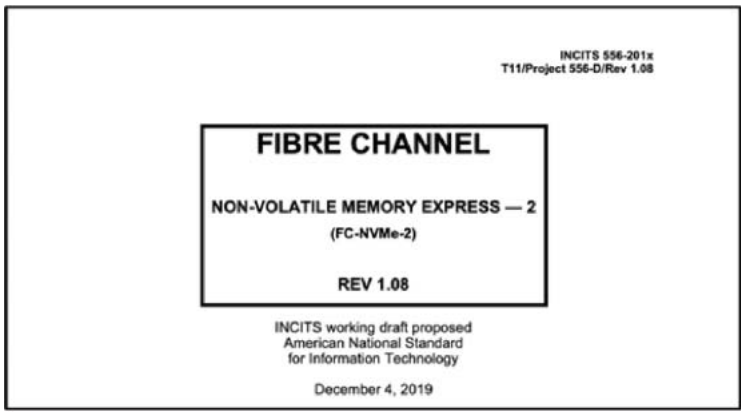
NVM Express Technical Proposal for New Feature

Technical Proposal ID	**TP 4028**
Change Date	**01/09/2018**
Builds on Specification	**NVM Express 1.3**

Technical Proposal Author(s)

Name	Company
Fred Knight	NetApp
David Black	Dell EMC
Sagi Grimberg	LightBits Labs

図 2.10　FC 標準を使用した NVMe コマンドおよびデータ転送を定義する NCITS FC-NVMe-2 Rev 1.08 T11-2019-00210-v004


INCITS 556-201x
T11/Project 556-D/Rev 1.08

FIBRE CHANNEL

NON-VOLATILE MEMORY EXPRESS — 2
(FC-NVMe-2)

REV 1.08

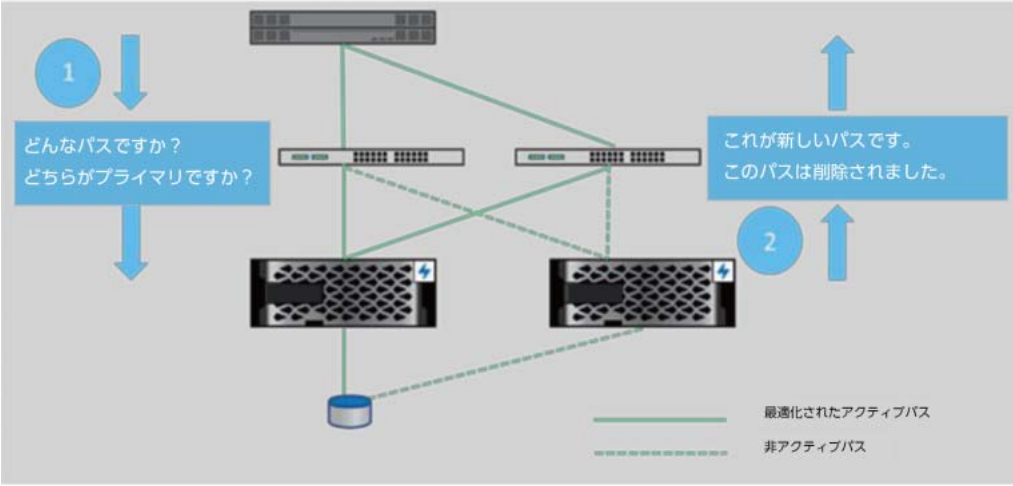
INCITS working draft proposed
American National Standard
for Information Technology

December 4, 2019

ANA には、以下の 2 つのコンポーネントがあります。

- イニシエータ側の ANA は、ターゲットに対してプライマリ対セカンダリなどのパス属性についてのクエリーを実行します。このデータは、パスを最適化するためにイニシエータ MPIO スタックによって使用されます。
- ターゲット側の ANA は、パス状態の変化を通知します。

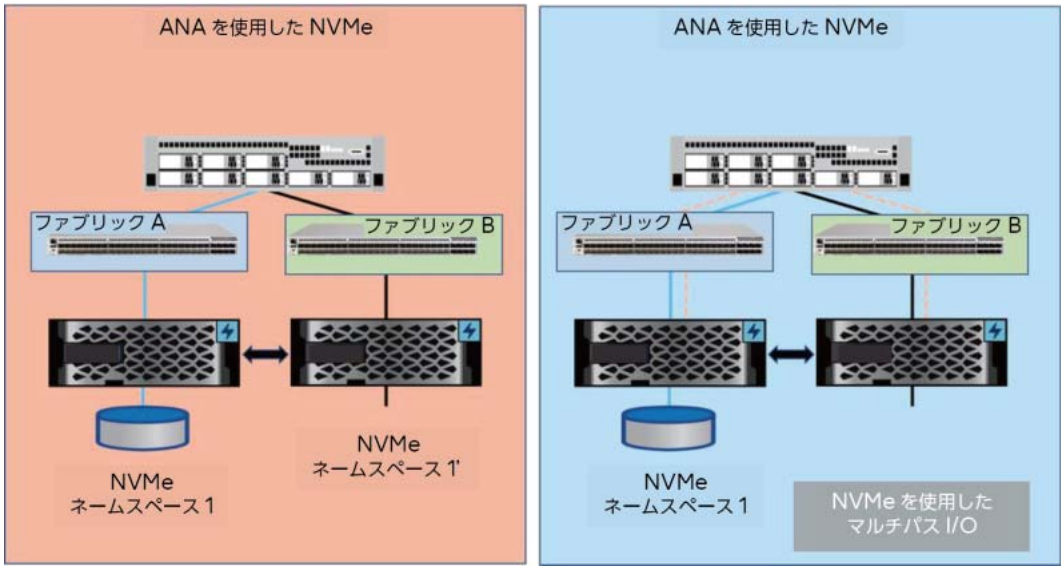
図 2.11　NVMe/FC ストレージフェイルオーバー：ANA を導入した ONTAP



ONTAP は、ALUA の実装に似たリモート I/O サポートを提供します。

図 2.12 に、ANA を使用した場合と使用しない場合の NVMe/FC の比較を示します。

図 2.12　ANA を使用した場合と使用しない場合の NVMe/FC の比較



2.3　NVMe over Fibre Channel

今後、データストレージシステムは NVMe にアップブレードされていきますが、どのような NVMe-oF 転送を選択すればよいのでしょうか。

RDMA トランスポートは重要ですが、最初は NVMe over Fibre Channel (NVMe/FC) がデータセンターファブリックでの主要な転送方法となる可能性があります。FC を転送に使用すると、以下のようなメリットがあります。

- 現在、ほとんどすべての高性能なブロックワークロードが FCP で実行されています。
- これらの組織のほぼすべて (約 70%) が、現在 FCP 対応の SAN を使用しています。
- 現在、性能を重視するほとんどのワークロードでは、ファブリックに第 5 世代または第 6 世代 (16Gbps または 32Gbps) のスイッチがすでに組み込まれています。
- 現在データセンターに設置されている 25/50/100 Gbps の Ethernet スイッチは、RDMA over IP、TCP、RoCE、またはその他の同様の転送用のバックボーンインフラストラクチャを形成するため、設置面積が小さくなっています。
- FCP と NVMe/FC は、SCSI-3 と NVMe を同時に転送するために同じ物理コンポーネントを使用できます。
- お客様の多くは、NVMe/FC の実行に必要なすべてのハードウェアをすでに所有しており、ONTAP へのシンプルなソフトウェアアップグレードで NVMe/FC の使用を開始できます。

FCP と NVMe は、共通のハードウェアコンポーネントとファブリックコンポーネントをすべて共有でき、同じケーブル (技術的には光ファイバ) 、ポート、スイッチ、ストレージコントローラー上で共存できます。このように、組織は自分のペースで NVMe に移行できるため、NVMe に容易に移行できます。事実、スイッチとダイレクタをアップグレードしたばかりの場合 (スイッチとダイレクタの世代が 5 または 6 である場合)、無停止で ONTAP にアップグレードできます。

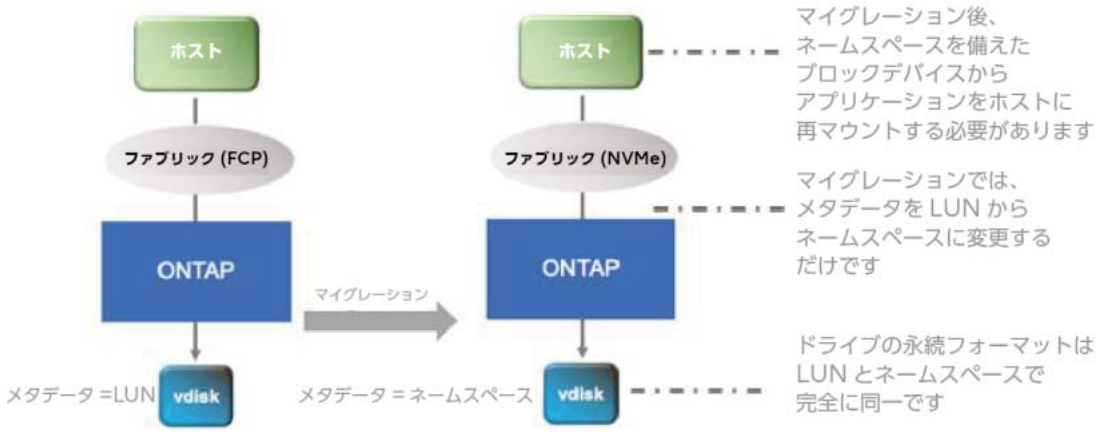
NVMe/FC は、FC フレーム内に SCSI-3 CDB をカプセル化するよう定義された FCP と非常によく似ています。NVMe/FC は古い SCSI-3 CDB を新しく合理化された NVMe コマンドセットに置き換えています。この単純な交換により、NVMe/FC はスループットとレイテンシを大幅に向上させます。

最初に NVMe/FC が発表されたのは、これが SAN プロトコルの主流であり、2 番目に規模が大きいプロトロコルである iSCSI の約 3 倍のシェアがあるためです。これは、組織がすでに重要な FC インフラストラクチャとスキルセットへの投資を行っていることを意味します。さらに、性能が最重要課題である場合は、FC SAN がほぼ常に最適な転送手段となります。

NVMe/FC は単にコマンドセットを SCSI から NVMe に交換するだけなので、移行は容易です。NVMe/FC は同じ FC 転送を使用するため、ホストからスイッチを経由してストレージアレイの NVMe/FC ターゲットポートに至るまで、同じハードウェアを使用します。したがって、NVMe/FC の実装では、HBA、スイッチ、ゾーン、ターゲット、ケーブルなどの既存の FC インフラストラクチャを使用できます。

ONTAP は FCP LIF とは別の NVMe/FC LIF を使用しますが、両方の LIF をホストイニシエータとストレージターゲットの両方の同じ物理 HBA ポートでホストすることができます。NVMe/FC と FCP は同じ物理インフラストラクチャを同時に共有できるため、同じ物理ポート、ケーブル、スイッチポート、およびターゲットポートが同時に FCP フレームと NVMe/FC フレームの両方をホストおよび送信することができます。2 つのプロトコルは物理層ではなく論理層で分離されているため、FCP から NVMe/FC への導入と移行はシンプルかつシームレスに行われます。本番運用を中断したり、複数の並列インフラストラクチャを実行したりすることなく、自分のペースで FCP から NVMe/FC にワークロードを移行できます。NVMe/FC と FCP は同じ物理インフラストラクチャを使用するため、お客様は、図 2.13 に示すように、新しいテクノロジーを無停止で実装してパフォーマンスを向上させ、新しいワークフローを導入し、既存ワークフローを移行することでパフォーマンスの向上を実現できます。

図 2.13　最新テクノロジーを無停止で実装



2.3.1　NVMe/FC リリースの発表

ONTAP は、業界初の NVMe/FC を導入しました。

NVMe/FC は、SCSI の代わりに NVMe コマンドセットを使用して、FCP や iSCSI などのブロックをホストに提供する新しいブロックアクセスプロトコルです。NVMe アーキテクチャは、効率的なコマンドセットとスケーラブルセッションを構築することで、レイテンシの大幅な削減と並列処理の増加を可能にし、メモリ内データベースや分析などの低レイテンシで高スループットのアプリケーションに適しています。

NVMe/FC は、オンボックスの ONTAP System Manager ソフトウェア (クラスタ管理ポートまたは任意のノード管理ポートの IP アドレスを Web ブラウザで指定して使用) または CLI を使用して、プロビジョニングおよび設定できます。

この新しいプロトコルを使用して最大のパフォーマンスを得るには、ホストから SAN ファブリック経由で ETERNUS AX series へのエンドツーエンドの NVMe/FC 接続が必要です。

備 考

- NVMe/FC の ONTAP 実装には、アプリケーションレベルの高可用性が必要です。コントローラーの損失またはパスの障害が発生した場合、アプリケーションホストはその (アプリケーションの) HA パートナーへのパスフェイルオーバーを実行する必要があります。この制限が存在するのは、SCSI プロトコルの ALUA に類似した、ANA と呼ばれる NVMe マルチパス仕様がまだ開発中であるためです。
  NVMe/FC の実装中に、NVMe フォーラムにおいて ANA プロトコルの設計を支援しました。NVMe フォーラムでは、最近、技術提案が認証されています。ONTAP の将来のリリースでは、この拡張機能がサポートされる予定です。
- 富士通のドキュメントや UI には、NVMe over Fiber Channel を、現在標準で商標登録されている NVMe/FC ではなく FC-NVMe と記載しているものがあります。FC-NVMe と NVMe/FC は代替可能な略称であり、どちらも NVMe over Fibre Channel を指します。

2.4 NVMe over TCP (NVMe/TCP)

ONTAP 9.10.1 では、ONTAP 初の Ethernet ベースのプロトコルである NVMe/TCP を追加しました。NVMe/TCP は、ONTAP が 2 番目にサポートする NVMe ブロックプロトコルとして NVMe/FC に加わります。NVMe/TCP は、Ethernet 上で Transmission Control Protocol (TCP) を使用します。転送用の TCP と Ethernet の組み合わせにより、NVMe/TCP は TCP と Ethernet が存在する任意の場所に実装できます。所有するデータセンターにほぼ限定される FC とは異なり、TCP と Ethernet はどこにでも存在可能です。さらに、NVMe/TCP には実際のネットワークハードウェアの制限はありません。ほとんどすべての Ethernet ネットワーク機器でサポートされています。たとえば、1 Mbps のネットワークインターフェースカード (NIC) と 1 Mbps のスイッチやルータで NVMe/TCP を実行できます。パフォーマンスに不満があるかもしれませんが、このソリューションはうまくいきます。ポイントは、実際のハードウェア要件がなく、Ethernet と TCP のほぼ普遍的な性質であるため、NVMe/TCP はほとんどどこでも実行できるということです。これは、クラウドの急速な成長において特に重要です。NVMe/TCP は、企業のデータセンター、サードパーティのホスティング、およびさまざまなクラウドエンドポイント間を接続できる SAN またはブロックプロトコルとして iSCSI に加わります。NVMe/TCP の最大の強みは、FC や iSCSI などの SCSI ベースのプロトコルから NVMe/FC や NVMe/TCP などの NVMe ベースのプロトコルに移行したときに得られる効率性、TCP と Ethernet の柔軟性と移植性です。本書の執筆時点 (2023 年 1 月) では、Red Hat Enterprise Linux、SUSE Enterprise Linux、Oracle Linux、VMware ESXi が NVMe/TCP をサポートしています。今後、他のオペレーティングシステムでもサポートが追加される予定であるため、現在 NVMe/TCP をサポートしているオペレーティングシステムのリストについては、富士通の担当営業に必ず確認してください。

備 考

SCSI と同様、NVMe プロトコルは転送に依存しません。つまり、NVMe ネームスペースにアクセスできるということになります。ネームスペースには、1 つ以上の NVMe-oF プロトコルを個別に、または同時に使用します。

第 3 章
NVMe の導入

NVMe-oF の初期導入の大半は、本番ワークロードを NVMe/FC や NVMe/TCP に移行する前に、組織内でNVMe/FCやNVMe/TCPの検証と認証をしているところから行われると予測しています。大部分の企業のストレージチームは、リスクを回避するため、新しいプロトコルを本番環境に導入する前に、徹底的な認証と検証を行いたいと考えています。また、初期に導入する企業のほとんどが、ANA が ONTAP NVMe/FC ターゲットに追加され、希望するホスト OS が NVMe/FC サポートの一部として ANA をサポートするまで待つと予想しています。採用に当たって ANA が未サポートでも関係がないのは、ストレージ層ではなくアプリケーション層で高可用性を管理するアプリケーションを持っている人たちだけです。前述のように、これらのアプリケーションの一部には MongoDB または Oracle ASM が含まれる場合があります。

3.1 NVMe/FC と NVMe/TCP のどちらを展開するかを選択するタイミング

ほとんどの場合、新規の NVMe-oF がグリーンフィールドで展開されることはまれです。既存のデータセンター運用への追加が大部分を占めています。つまり、ある転送方式が別の転送方式よりも有利になるような既存のインフラが存在する可能性が高いということです。もちろん、すでに導入されているさまざまなワークフローやインフラの要件に基づいて転送方式を組み合わせたり、一致させたりできないという要件はありません。

最初に理解すべきことは、両方の NVMe-oF 転送方式が、iSCSI や FCP などの SCSI ベースのブロックプロトコルと非常によく共存していることです。さらに、NVMe/FC は FCP とまったく同じコンポーネントを同時に使用できますが、iSCSI と NVMe/TCP という 2 つの Ethernet プロトコルについても同じことが言えます。実際、FCP と iSCSI のいずれかまたは両方を使用した十数台の LUN や NVMe/FC と NVMe/TCP のいずれかまたは両方を使用した別の十数台のネームスペースに同じホストからアクセスすることが、まったく同じ物理 HBA と NIC のポート、ケーブル、スイッチ、および ONTAP コントローラーを利用してとても簡単に実施できます。これにより、4 つのブロックプロトコル間の共存と移行が非常に容易になります。

NVMe-oF 転送方式を特定のワークフローで展開することを検討する場合、問題は何を最適化するかです。所有しているデータセンター内で完全なパフォーマンスを求めている場合は、おそらく NVMe/FC が最初の選択肢となります。柔軟性と移植性を最適化しようとする場合、特に FC ファブリックが存在しない場合や、所有しているデータセンターとクラウドエンドポイントを接続する必要がある場合は、NVMe/TCP が最初の選択肢になる可能性があります。

3.2　ONTAP 機能のサポートと共存

表 3.2 および表 3.3 では、ONTAP ツールと機能のうち、現在サポートされているもの、共存可能なもの、NVMe-oF に対応していないものを記載しています。これらの表のすべての項目は、本書の執筆時点の情報です。ONTAP NVMe-oF ターゲットの機能とパフォーマンスは向上し続けているため、未サポートの表に記載された機能の多くが、時間の経過とともにサポートされる可能性があります。

表 3.1　SCSI および NVMe の用語

FCP/iSCSI	NVMe-oF の備考	備考
FCP - ワールドワイドポート名 (WWPN) iSCSI-iSCSI 修飾名 (IQN)	NVMe 修飾名 (NQN)	一意の識別子
SCSI ターゲット	NVMe サブシステム	ストレージオブジェクトとアクセスポイントを持つエンティティ
ポート	ポート	通信用アクセスポイント
I_T nexus	NVMe コントローラー (複数のキューペアを使用)	イニシエーターとターゲット間のセッション
LUN	ネームスペース	ストレージオブジェクト
Asymmetric logical unit access (ALUA)	ANA	非対称アクセス特性

表 3.2　NVMe によってサポートされている ONTAP 機能および NVMe と共存できる ONTAP 機能

ONTAP リリース	NVMe-oF 機能
9.7	• NVMe/FC • シングルノードのみ (高可用性なしなど) • 4k ブロックサイズ • Asymmetric Namespace Access (ANA) を備えた 2 ノード HA • NVMe-oF ライセンス • コピーと書き込み (CAW) • 512b ブロックサイズ • 読み取り専用のネームスペース • 変更された NS リスト / ベンダー固有のログページなどの追加のログページ • NVM-oF 同期 SnapMirror
9.8	• 同一 SVM 内での LUN とネームスペースの共存
9.9	• ETERNUS AX series ASA に対応する NVMe/FC • ETERNUS AX series ASA の大きいネームスペース (最大 128 TB まで拡張) • NVMe/FC VMware vSphere 仮想ボリューム (vVol) • NVMe-oF 中断 • 4 ノードアレイ上の NVMe/FC

ONTAP リリース	NVMe-oF 機能
9.10.1	• NVMe/TCP • ネームスペースのサイズ変更 • NVMe-oF キャンセル • ETERNUS HX series での NVMe-oF • NVMe/TCP 上の IPSec
9.11.1	• NVMe/TCP パフォーマンスの強化 • LUN ⇔ ネームスペースの双方向インプレース変換 • NVMe-oF のスケール拡張 (最大 12 ノードのサポートなど)
9.12.0	• Cloud Volumes ONTAP (アマゾンウェブサービス [AWS] および Azure) および FSx での NVMe/TCP
9.12.1	• GCP Cloud Volumes ONTAP での NVMe/TCP • NVMe-oF インバンド認証 • MCC IP 上の NVMe/FC • NDO を使用した NVMe-oF の SnapMirror Sync

表 3.3　NVMe で現在サポートされていない ONTAP 機能

機能	メモ
ETERNUS AX series ASA (All SAN Array) 上の NVMe-oF Symmetric Active/Active アクセス	ETERNUS AX series ASA では FCP と iSCSI だけが A/A であり、NVMe-oF では対応していません
ネームスペース (NS) のサービス品質 (QoS)	QoS はボリュームおよび SVM レベルでのみサポートされ、NS レベルではサポートされません
ネームスペースの移動	許可されていません (強制)
NVMe/TCP vVol	未サポート (現在サポートされているのは NVMe/FC vVol のみです)
MCC IP 上の NVMe/TCP	未サポート (現在、MCC IP 4 パックで NVMe/FC をサポートしています)
NVMe/TCP での TLS	未サポート
SnapMirror Business Continuity における NVMe-oF	現在、SnapMirror Business Continuity でサポートされているのは FCP と iSCSI のみです
SVM 移行	ブロックのサポートなし
外部 LUN インポート (FLI)	FLI は FCP を使用してすべての移行を実行します

備 考

インポートされた LUN は、インポートの完了後、組み込まれている双方向 (SCSI LUN および NVMe ネームスペース間) インプレース変換ユーティリティを使用して、ネームスペースに変換できます。LUN またはネームスペースに関するメタデータのみを変更するため、このユーティリティは非常に高速です。

3.3 相互運用性

SUSEおよびBroadcomとの提携により、本番環境に適用可能なNVMe/FCを市場に投入しました。現在サポートされている構成のリストについては、富士通の担当営業に確認してください。

3.3.1 富士通のサポート状況の確認方法

NVMe-oF は、以下の 3 つの異なる軸で急速に開発されているため、構成の確認が重要です。

- NVMe-oF ターゲット
- 関連するすべてのスイッチ、HBA など
- 現在 NVMe-oF をサポートする以下のホスト OS
  - Linux (RHEL、Oracle Linux、および SLES)
  - VMware ESXi
  - Windows

■ NVMe/FC のみ

- HBA (32G 以下、Gen6 以降) と一致するドライバ、ファームウェア
  - Broadcom (Emulex)
  - Marvell (Qlogic)
- FC スイッチ (16G 以下、Gen 5 以降) のスイッチ OS
  - Broadcom (Brocade)
  - Cisco

3.4 LUN のネームスペース変換、またはネームスペースの LUN 変換

ONTAP 9.11.1 では、組み込みの双方向（LUN およびネームスペース間）インプレース変換ユーティリティが追加されました。LUN とネームスペースの間の変換は、非常に高速 (メタデータのみ変更) で、非常に容易です。これにより、NVMe の採用が容易になり、SCSI と NVMe プロトコル間の移行がよりシームレスになります。

3.4.1 機能のハイライト

- 同じ ONTAP ボリューム内でのインプレース LUN ⇔ ネームスペース変換
- データコピーが不要。実際のユーザーデータは変更されず、メタデータの更新のみが必要です。
- 既存のスナップショットコピー (移行前に作成) へのアクセスが失われないため、データ管理と保護の継続が可能
- 変換時に識別子 (シリアル番号や UUID など) とともにブロックサイズを保持
- 変換後にパフォーマンスが低下しない (同様のプラットフォーム構成で新規作成した場合と比較)
- エンドツーエンドの LUN ⇔ ネームスペース変換プロセスを完了するために必要なホストを修復

3.4.2 機能制限

- 1 回のコマンドで LUN またはネームスペースを一括変換することは不可能
- マップされた LUN をネームスペースに変換することはできない (その逆も同様)
- FLI 関係にある LUN をネームスペースに変換することはできない
- SnapMirror Business Continuity 関係にある LUN をネームスペースに変換することはできない
- MetroCluster 構成の LUN をネームスペースに変換することはできない
- vVol バインディングを持つ LUN、または Protocol Endpoint (PE) として機能している LUN をネームスペースに変換することはできない
- 0 以外のプレフィックスおよび / またはサフィックスストリームを持つ LUN をネームスペースに変換することはできない (ONTAP はネームスペース用のプレフィックスおよびサフィックスストリームをサポートしないため)
- ブロックサイズが 4k のネームスペースを LUN に変換することはできない (ONTAP は 4k LUN をサポートしていないため)
- ネームスペースにサポートされた有効な `os_types` が設定された LUN のみ変換可能

表 3.4　LUN ⇔ ネームスペース変換ユーティリティ機能のサポート

LUN os_type	プレフィックスサイズ	サフィックスサイズ	LUN⇔NS が 可能かどうか
vmware	0	0	はい
hyper_v	0	0	はい
windows_2008	0	0	はい ('windows' に変換)
windows_gpt	17 (ONTAP 9.7) 0 (ONTAP 9.8 以降)	0	あり (プレフィックスが 0 の場合のみ 'windows' に変換)
windows	31.5 (ONTAP 9.7) 0 (ONTAP 9.8 以降)	0	はい (プレフィックスが 0 の場合のみ)
linux	0	0	はい
xen	0	0	はい
solaris	0	0	いいえ (サポートされていない NVMe os_type)
solaris_efi	17 (ONTAP 9.7) 0 (ONTAP 9.8 以降)	0	いいえ (サポートされていない NVMe os_type)
hpux	0	0	いいえ (サポートされていない NVMe os_type)
aix	0	0	いいえ (サポートされていない NVMe os_type)
openvms	0	0	いいえ (サポートされていない NVMe os_type)
image (デフォルト)	0	0	いいえ (サポートされていない NVMe os_type)

LUN とネームスペースの変換についての詳細な情報については、「付録 D LUN とネームスペース間の変換」(P.55) を参照してください。

第 4 章
NVMe/FC のベストプラクティス

4.1 NVMe/FC のベストプラクティス

組織が NVMe/FC ワークロードを検証して適格性を確認するか、実環境で使用するかにかかわらず、すべてのチームは一般的な FCP SAN のベストプラクティスに従う必要があります。NVMe/FC は FC を転送方式として使用するため、これらのベストプラクティスが適用されます。SAN のベストプラクティスについては、富士通マニュアルサイトの「ETERNUS AX/HX series ONTAP SAN 構成のベストプラクティス」を参照してください。

4.1.1 ファブリックとスイッチの構成と運用に関するベストプラクティス

NVMe/FC では、一般的な Brocade または Cisco FC スイッチおよびファブリックのベストプラクティスと異なる特別な構成やベストプラクティスは必要ありません。単一イニシエータゾーニングがベストプラクティスです。もう 1 つのベストプラクティスは、WWPN を使用して、スイッチポートベースのゾーンメンバーシップまたはハードゾーニングの代わりに、ゾーンメンバーシップを割り当てることです。

4.1.2 NVMe-oF のベストプラクティス：パス構成

インターフェースのシングルポイント障害を回避するために、SVM ごと、ノードごと、ファブリックごとに 2 つのパスをプロビジョニングすることを強く推奨します。ONTAP では、ストレージ管理者がノードごとまたは SVM ごとに 3 つ以上の LIF を作成することができません。これは、イニシエーターに提供されるパスの数を減らすために、SAN のターゲットエンジニアが LIF の作成を制限するという意識的な選択でした。この制限を設けることで、シングルポイント障害を排除するのに十分な冗長性を確保できると同時に、イニシエーターに提供するパスの数を制限するために SCSI ベースの構成で必要となる、選択的 LUN マップ (SLM) などの機能の必要性を軽減または排除できます。特定のノードに追加の LIF が必要な場合は、追加の SVM を作成してノードごとに追加の 2 つの LIF を作成し、ネームスペースとの I/O のバランスを取ることができます。

NVMe は、通信、アラート、パスの管理およびパス状態の変更管理をするための ANA プロトコルを追加しました。ANA は、以下の 2 つのコンポーネントで構成されています。

- 現在のパスの状態についての情報をターゲット (ONTAP ノード) に問い合わせる、ホスト側の実装。
- パスの状態が変化したときにアラートを生成し、使用可能なすべてのパスの一覧をイニシエータ側のクエリに応答するストレージノードの実装。

ホスト側の ANA 実装は、受信したすべてのパス情報をホストのマルチパススタックに渡す役割を担います。dm-multipath などのホストのマルチパススタックは、パスの優先順位と使用状況を管理します。

備 考

- ONTAP 9.9.1 では、NVMe-oF パスをアクティブ / 最適化 (AO) またはアクティブ / 非最適化 (ANO) としてアドバタイズパスに変更する NVMe-oF リモート I/O サポートが追加されました。これは、統合システム (ETERNUS AX/HX series) 上で SCSI ベースのプロトコルを使用するためです。
- SCSI プロトコルの iSCSI および FC は、ETERNUS AX series ASA 上で対称ですが、NVMe-oF プロトコルは、ETERNUS AX series ASA 上でも非 ASA 上と同じ非対称 AO/ANO 特性を持ちます。これは、ローカルパス操作とリモートパス操作に関する NVMe-oF プロトコルの違いによるものです。

このため、ONTAP ネームスペース専用の Linux NVMe マルチパスを使用することを推奨します。

4.1.3　マルチパスに関する推奨事項

ONTAP LUN 専用に Linux dm-multipath を使用し、ONTAP ネームスペースには NVMe マルチパスを使用することを推奨します。

以前の SUSE 15 (SLES 15) バージョンでは、NVMe マルチパスでラウンドロビン方式のロードバランシングを使用できなかったため、パフォーマンスが低下しました。この制限は、最近の SLES 15 カーネル /nvme-cli バージョンで解決されています。

4.2　NVMe-oF の設定と構成

4.2.1　設定および構成のクイックリスト

NVMe/FC や NVMe/TCP を設定する前に、以下を実施してください。

手順 ▶▶▶

1. 現在の構成が、サポートされている構成であることを確認します。サポートされていない構成の場合、ストレージの実装が最適ではなく、構成が不十分になるおそれがあります。
2. 富士通およびスイッチベンダーが出している SAN のベストプラクティスに従うように、物理インフラストラクチャの導入、ケーブル接続、および構成を実行します。
3. (NVMe/FC のみ) すべてのファブリックスイッチで N_Port ID 仮想化 (NPIV) を有効にします。
4. (NVMe/FC のみ) 単一イニシエーターゾーニングを使用し、WWPN を使用してゾーンのメンバーシップを指定します。スイッチポート接続を使用してゾーンのメンバーシップまたはハードゾーニングを指定しないでください。

5. ONTAP System Manager または ONTAP CLI を使用して、NVMe-oF オブジェクト (SVM、ボリューム、ネームスペース、サブシステム、および LIF) を作成します。詳細は、「付録 A ONTAP System Manager を使用した ONTAP NVMe/FC および NVMe/TCP オブジェクトの作成」および「付録 B ONTAP NVMe/FC および NVMe/TCP CLI コマンド：初期設定と検出」を参照してください。
6. Active IQ Unified Manager を使用して、新しく作成された NVMe オブジェクトの健全性とパフォーマンスを監視し、レポートのしきい値とアラートを作成します。

4.3 設定および構成手順の詳細資料

NVMe のセットアップおよび構成タスクの実行の詳細については、以下の資料を参照してください。

- ONTAP System Manager (オンボックス UI) を使用して NVMe/FC オブジェクトを ONTAP に設定および構成する場合は、「付録 A ONTAP System Manager を使用した ONTAP NVMe/FC および NVMe/TCP オブジェクトの作成」(P.45) を参照してください。
- CLI を使用して NVMe/FC を ONTAP に設定および構成する場合は、「付録 B ONTAP NVMe/FC および NVMe/TCP CLI コマンド：初期設定と検出」(P.50) を参照してください。
- NVMe オブジェクトを表示して I/O を実行する場合は、「付録 B ONTAP NVMe/FC および NVMe/TCP CLI コマンド：初期設定と検出」(P.50) を参照してください。

第 5 章
パフォーマンス

NVM Express 社は、フラッシュによって引き起こされるボトルネックを解決するために、最初に NVMe の仕様とアーキテクチャを開発しました。フラッシュでハードドライブを交換することで、ハードドライブによって引き起こされた主要なストレージのボトルネックは解消されましたが、コマンドセットとコントロールプレーンに別のボトルネックが発生しました。NVMe の仕様とアーキテクチャは、SCSI コマンドセットを以下のような新しく合理的なコマンドセットに置き換えます。

- コマンドの合理化 (SCSI コマンドは、約 40 年前に最初に作成された標準版の下位互換版がありました)
- ハードウェア割り込みの代わりにポーリングモードを使用
- コンテキストスイッチの削減
- ロックなし
- キューを 64k (65,535) まで拡大。各キューの深度は 64k とする。

図 5.1 は、上記の各ポイントがスループットとレイテンシに与える影響を示しています。合理化したコマンドセット (I/O パス) の方が効率的であり、同じワーキングセットとインフラを使用して、同じ時間でより多くの I/O を実行できます。コンテキストスイッチを削減し、ハードウェア割り込みではなくポーリングモードを使用することで、NVMe は強制的なプロセッサ待機時間を大幅に削減します。ソフトウェアロックを解除すると、プロセッサの待機時間も短縮されます。パフォーマンスの最大の向上は、キューの数が非常に多いことと、キューに関連付けられたキューの深さで実現可能です。各キューは別々のプロセッサコアを使用して同時に I/O を処理できます。

これらの機能強化により、パフォーマンスが大幅に向上します。こういったパフォーマンス向上は、スループットまたは IOPS の増加とレイテンシの減少という形で真っ先に現れます。最初の検証では、ワークロードエンジニアリングチームとパフォーマンスチームは、レイテンシを 80 ～ 100µs 削減しながら IOPS を測定した結果、時に 50% を超えるパフォーマンスの向上を確認しました。

図 5.1　NVMe/FC の超高速性能設計

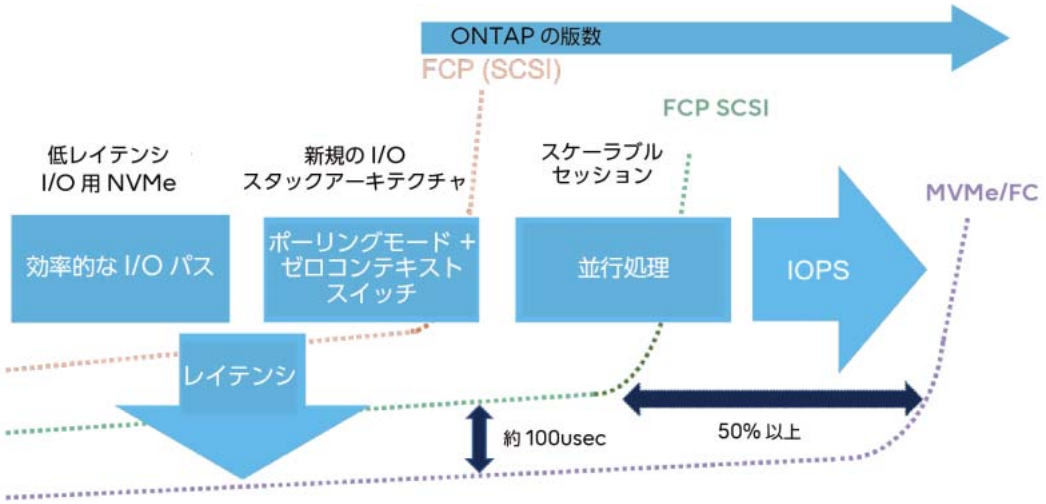


表 5.1 に、内部テストの結果の一部を示します。単一の LUN と単一のネームスペースのアクセスを比較すると、大幅に増加しています。この比較は、並列化の増加がパフォーマンスにどの程度影響するかを示しています。ストリームの I/O 処理が単一の CPU コアに制限されるため、パフォーマンスが必要な場合に単一の LUN を使用することは推奨していません。

表 5.1　4k ランダムリード時の NVMe/FC と FCP の比較

	NVMe/FC	FCP との差分 (割合)
シングルポート、IOPS	619K	207%
単一のネームスペース /LUN、IOPS	540K	880%
最大 IOPS	865K	+51%

図 5.2 および図 5.3 では、弊社の性能特性調査部門による 8k および 4k のランダムリード特性について示します。試験は、いくつかの一般的なクリティカルエンタープライズアプリケーションについて弊社、Brocade、Broadcom が検証したリファレンスアーキテクチャを定義、構築、試験、文書化することによって行われました。

図 5.2　高可用性 (HA) ペア、8k ランダムリード時の FCP と NVMe/FC の比較結果

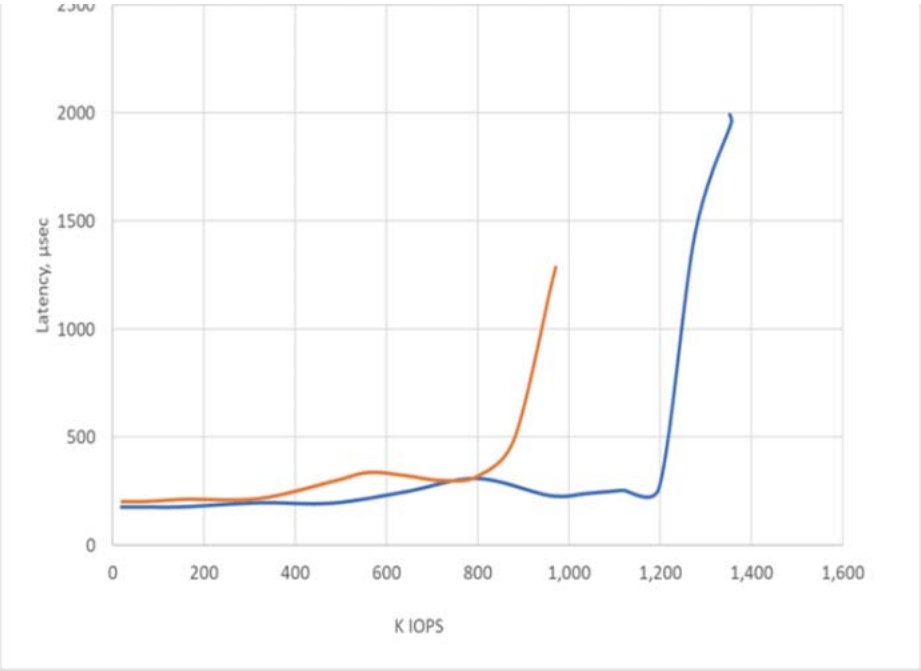


図 5.3　HA ペア、4k ランダムリード時の FCP と NVMe/FC の比較

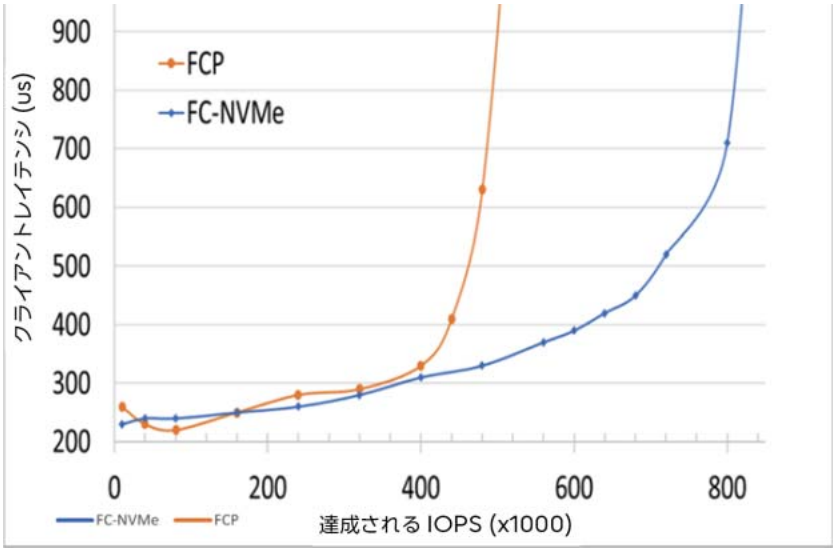
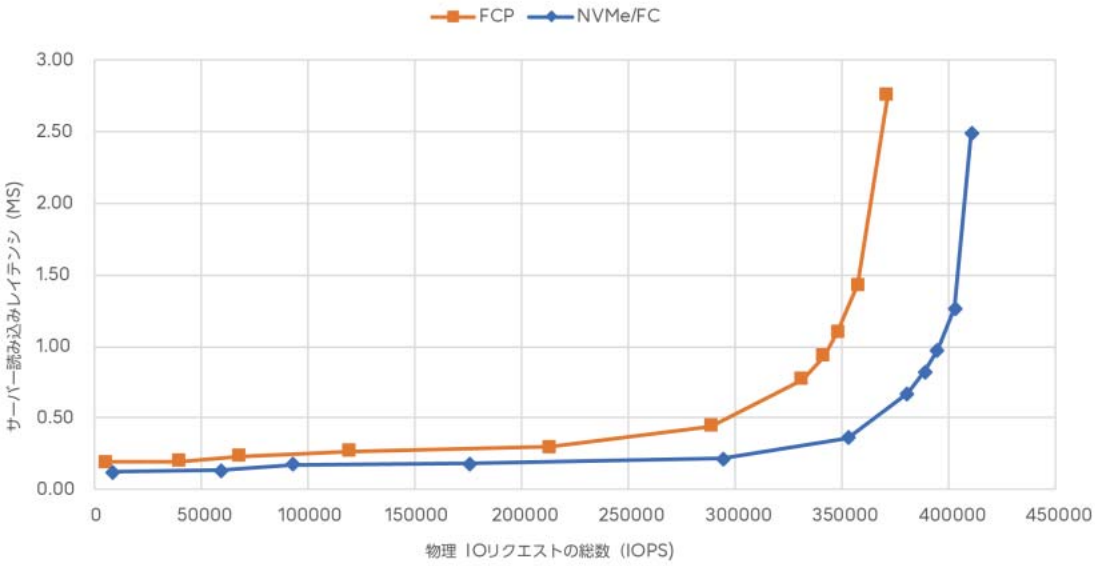


図 5.4 では、FCP と NVMe の両方について、レイテンシあたりの IOPS 数を比較したものです。これらは、コントローラーが特定の期間内に完了できる I/O 数の大幅な増加を示しています。興味深いことに、NVMe/FC を使用すると、達成される IOPS の数が増加し、IOPS の完了に必要な時間が短縮されます。

図 5.4　FCP から NVMe/FC への移行によるパフォーマンスの向上



第 6 章
NVMe/TCP のベストプラクティス

NVMe/TCP は Ethernet と TCP/IP を使用するため、ネットワークパフォーマンスを向上させるものは、iSCSI の場合と同様に、NVMe/TCP のパフォーマンスにプラスの影響を与える可能性があります。一般的に、スイッチおよびルータに関するネットワークベンダーのベストプラクティスに従うことを推奨します。

第 7 章
NVMe-oF の機能拡張

7.1 ONTAP 9.7

7.1.1 512 バイトブロック

NVMe のブロックサイズは、すべての OS で 4096 または 4k です。ブロックサイズを 512 バイトブロックに減らすことで複数の ESX の 512 バイトを 4k の ONTAP ブロックにアグリゲートする必要がある代わりに、共通のブロックサイズを提供して ESXi との相互運用がより簡単になっています。512 バイトブロックのサポートにより、ONTAP で ESXi のコピーおよび書き込みや、Atomic Test and Set (ATS) をサポートする機能も強化されています。

7.2 ONTAP 9.9.1

7.2.1 NVMe-oF リモート I/O サポート

NVMe-oF にリモート I/O サポートが追加されました。これにより、NVMe-oF パスがアクティブ / 非アクティブモデルから、他のすべての ONTAP ブロックプロトコルが使用するアクティブ最適化 (AO)/ アクティブ非最適化 (ANO) モデルに変更されます。

図 7.1 は、リモート I/O をサポートしない NVMe-oF、図 7.2、リモート I/O をサポートする NVMe-oF です。

図 7.1　リモート I/O をサポートしない NVMe-oF

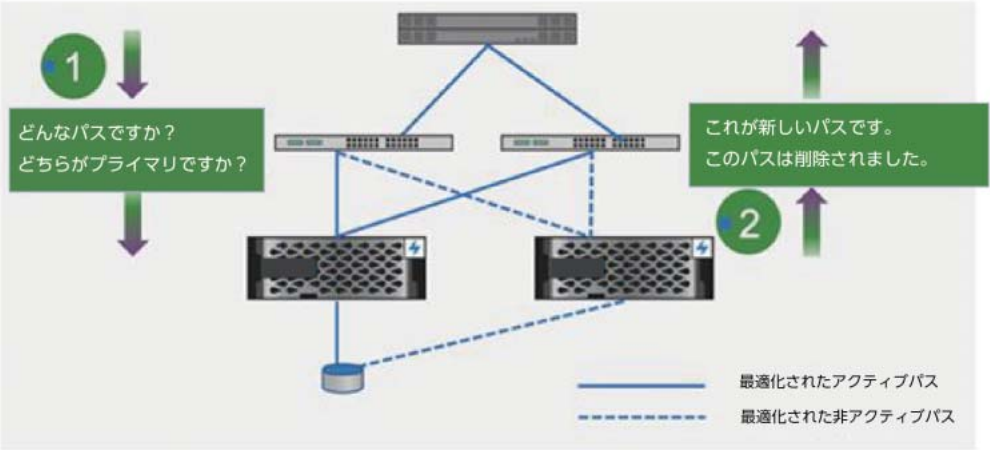


図 7.2　リモート I/O をサポートする NVMe-oF

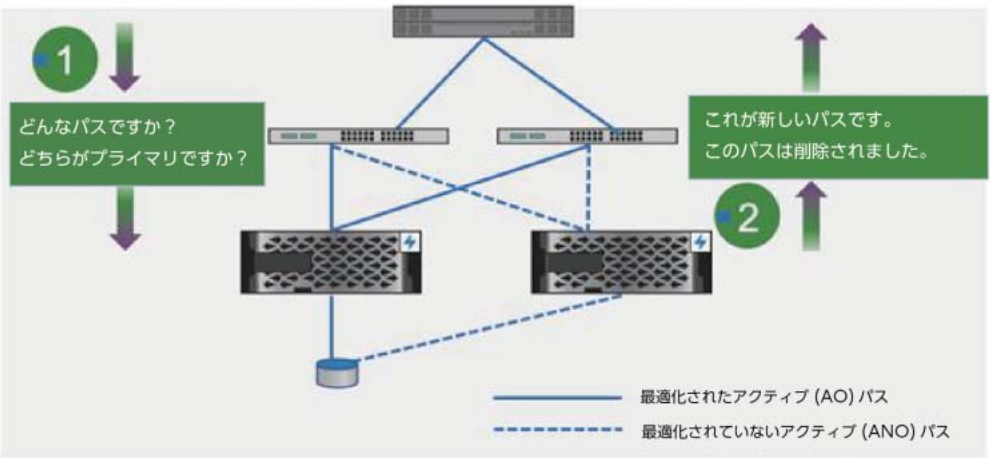


図 7.1 と図 7.2 を比較すると、大きな違いはないように見えますが、ごく小さな違いがあります。リモート I/O では、NVMe-oF でのすべてのパスのサポートがアクティブになります。つまり、これらのパスのいずれかに送信された I/O が確認され、反応または応答されます。以前は、リモート I/O がなかったため、非アクティブなパスが無効となり使用できませんでした。

```
All SAN Array adds NVMe/FC support
ONTAP 9.9.1 adds NVMe/FC as an additional block protocol. Unlike either FC or iSCSI, NVMe/FC on
All SAN Array (ASA) will continue to be asymmetric (AO/ANO). This assignment is due to
differences in how NVMe-oF works with remote versus local paths.
```

7.3　ONTAP 9.10.1

7.3.1　NVMe/TCP の導入

ONTAP 初の Ethernet ベースにして第 2 の NVMe-oF プロトコル転送方式です。

7.3.2　ネームスペースのサイズ変更

ネームスペースのサイズを変更する機能が導入されました。

7.3.3　大規模ネームスペース

ETERNUS AX series ASA で、ネームスペースの最大サイズが 128 TB に増加しました。大規模なネームスペースのサポートは、ONTAP 9.12.1 のパブリックプレビューで ETERNUS AX/HX のコントローラーに追加され、ONTAP の次のリリースで一般に利用可能な機能になります。

7.3.4 検出コントローラーでの非同期イベント要求 (AER、オペコード OC) のサポート

非同期イベント要求 (AER) は、新しいマップまたはマップを持つサブシステムが追加されたときにトリガーされます。AER は、マップが追加されたサブシステムの一部であるすべてのアクティブセッションに発行されます。アンマップのサポートは、ONTAP の将来のリリースで追加される予定です。AER は、他の admin キュー要求と同様に処理されます。AER の完了には時間がかかる場合があります。

7.4 ONTAP 9.11.1

7.4.1 NVMe/TCP パフォーマンスの拡張

ONTAP 9.11.1 では、以下の NVMe/TCP パフォーマンスの拡張が導入されました。

- 書き込み用のカプセル化データのファーストバースト
- 読み取りに対する読み取り応答時間の縮小

これらの機能拡張により、NVMe/TCP の性能が大幅に改善し、複数の I/O ワークロードにおいては iSCSI と同等か iSCSI を超える性能を実現しました。

7.4.2 LUN からネームスペースへの双方向変換ユーティリティ

非常に高速で使いやすい、LUN からネームスペースへの変換、およびネームスペースから LUN への変換ユーティリティです。このユーティリティは、ONTAP に組み込まれています。LUN/ ネームスペースに関するメタデータのみを変更するインプレース変換であるため、変換は非常に高速です。

LUN とネームスペースの変換についての詳細な情報については、「付録 D LUN とネームスペース間の変換」(P.55) を参照してください。

7.5 ONTAP 9.12.0

7.5.1 AWS FSx と Cloud Volumes ONTAP で導入された NVMe/TCP のサポート

- FSx HA に NVMe/TCP サポートが追加されました。
- Cloud Volumes ONTAP AWS HA に NVMe/TCP サポートが追加されました。
- Cloud Volumes ONTAP Azure HA に NVMe/TCP サポートを追加されました。

7.6 ONTAP 9.12.1

7.6.1 NVMe/TCP によるその他のクラウドサービスのサポート

- FSx シングルノードに NVMe/TCP サポートが追加されました。
- Cloud Volumes ONTAP AWS Single Node に NVMe/TCP のサポートが追加されました。
- Cloud Volumes ONTAP Azure シングルノードに NVMe/TCP サポートが追加されました。
- Cloud Volumes ONTAP GCP シングルノードおよび HA が実装されました。

7.6.2 NVMe/TCP に対する双方向インバンド認証

ONTAP 9.12.1 以降では、NVMe ホストとコントローラー間のセキュアな双方向認証が、DH-HMAC-CHAP 認証プロトコルを使用する NVME-TCP でサポートされます。

各ホストまたはコントローラーは、NVMe ホストまたはコントローラーの NQN と、管理者が NVMe ホストまたはコントローラーのピアを認証するために設定した認証シークレットを組み合わせた DH-HMAC-CHAP キーに関連付ける必要があります。このキーは、ピアに関連付けられたキーを認識している必要があります。SHA-256 はデフォルトのハッシュ関数です。

7.6.2.1 復元する前に確認すること

NVMe/TCP プロトコルを実行していて、DH-HMAC-CHAP を使用したセキュア認証を確立した場合は、復元をする前に DH-HMAC-CHAP を使用したホストを NVMe サブシステムから削除する必要があります。ホストが削除されない場合、復元は失敗します。

7.6.3 MCC IP に対する NVMe/FC のサポート

ONTAP 9.12.1 では、MCC IP に NVMe/FC サポートが導入されました。詳細は、「付録 F MCC IP での NVMe/FC の構成と設定」(P.60) を参照してください。

付録 A
ONTAP System Manager を使用した ONTAP NVMe/FC および NVMe/TCP オブジェクトの作成

ONTAP System Manager を使用して NVMe オブジェクトを作成するには、以下の手順を実行します。

手順 ▶▶▶

1 NVMe をサポートする SVM を作成します。

備 考

この手順では、このワークフローの残りの部分で作成されたすべての NVMe ストレージオブジェクトを格納する SVM を作成します。

1-1 ONTAP System Manager で、[Storage] > [SVM] に移動します。[Create] をクリックします。

1-2 NVMe を選択すると、SVM セットアップダイアログボックスの一部として、NVMe を構成するサブシステム、NVMe Qualified Name (NQN) 、およびネームスペースの情報を作成または定義するプロンプトが起動します。[Submit&Continue] をクリックします。

備 考

- ホスト NQN を表示するには、以下の Linux コマンドを実行します。

```
# cat /etc/nvme/hostnqn
```

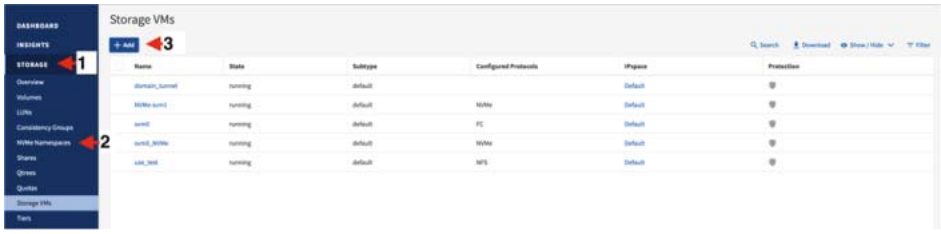
- SVM の NQN を表示するには、以下のコマンドを実行します。

```
vserver nvme show -vserver <vserver_name>
tme-a800::> nvme show -vserver NVMe-svm1
  (vserver nvme show)

            Vserver Name: NVMe-svm1
  Administrative Status: up
Discovery Subsystem NQN: nqn.1992-08.com.netapp:sn.70c0b1366f3611edacf100a098e22473:discovery
```

- [Storage] > [NVMe] > [NVMe Namespaces] に移動して、ONTAP System Manager でサブシステムの NQN を表示することもできます。次に、NQN を表示するネームスペースへのリンクをクリックします。

図 A.1　ONTAP System Manager - SVM の作成



1-3 以下のいずれかを実行します。

- SVM 管理者ダイアログボックスで、SVM 管理者の詳細を設定します。
- 特定の SVM 管理アカウントの追加を省略するには、[Skip] をクリックします。

図 A.2　ONTAP System Manager - SVM の作成：NVMe 転送方式の構成 - NVMe/FC および NVMe/TCP

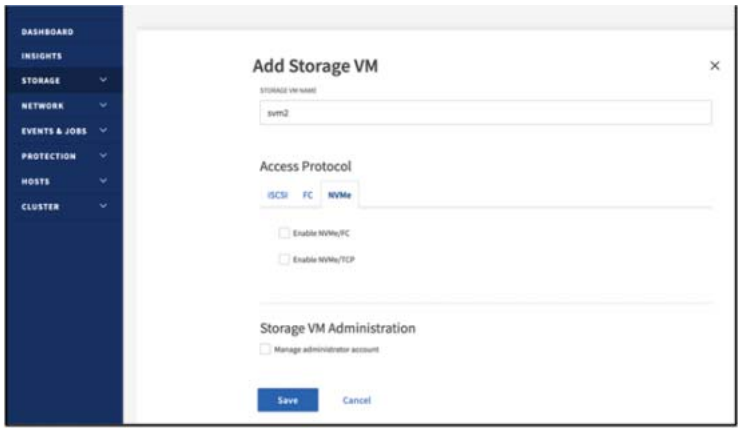


図 A.3　ONTAP System Manager - SVM の作成：NVMe/FC の構成

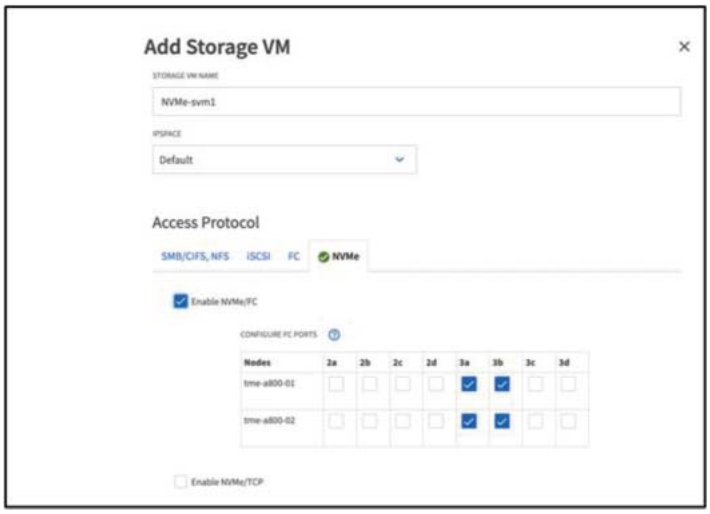


図 A.4　ONTAP System Manager - SVM の作成：NVMe/TCP の構成

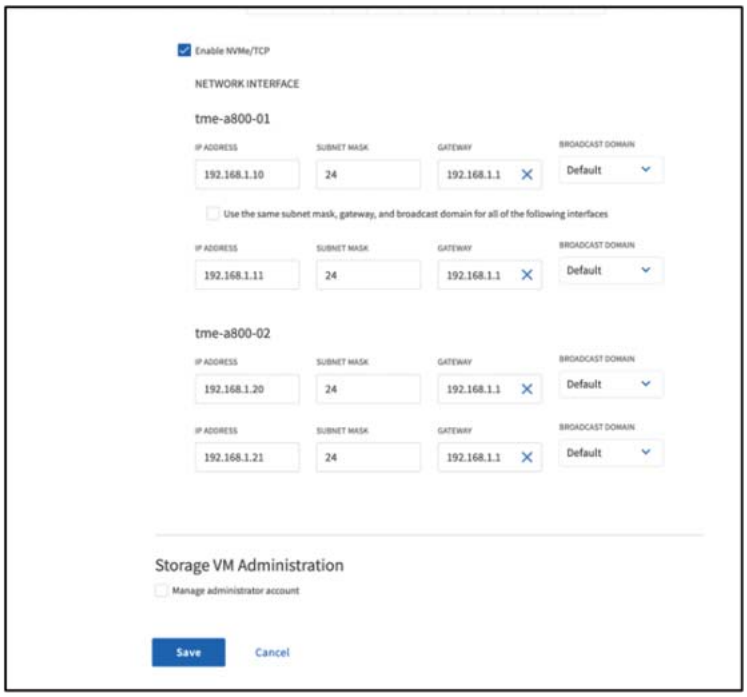
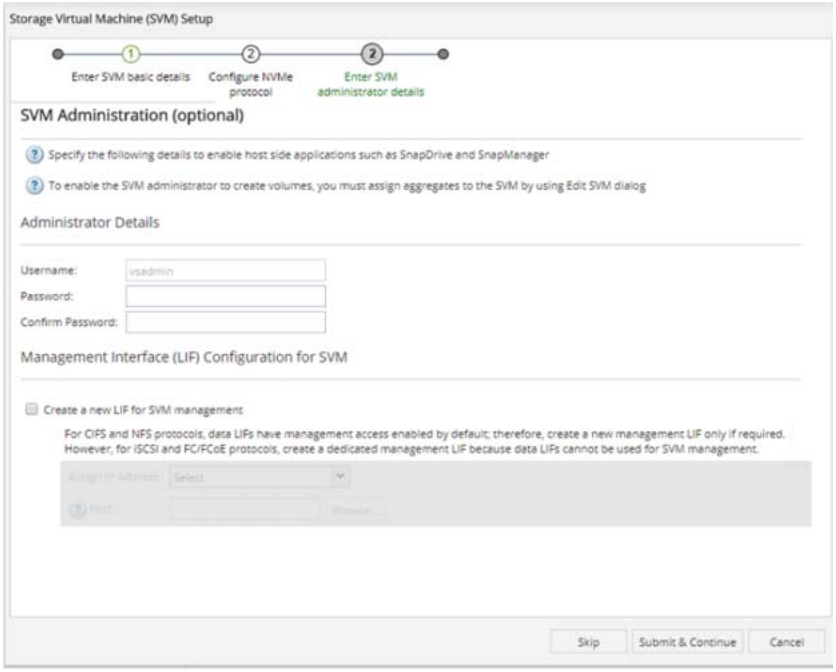


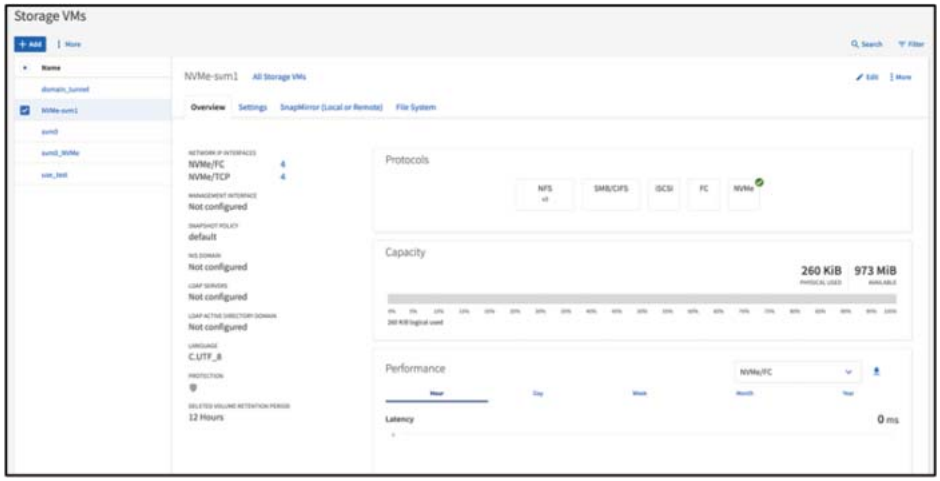
図 A.5　ONTAP System Manager - SVM の作成：管理者の詳細設定



2　作成された SVM の概要を確認し、[OK] をクリックします。

3 新しく作成した SVM を選択します。すべてのプロトコル設定とサービスステータスを確認するには、トップメニューから [SVM Settings] をクリックします。

図 A.6　新しく作成した SVM の表示



4 [SVM Dashboard] ページに戻るには、SVM 設定ページの右上隅にある [Back] をクリックします。[SVM Dashboard] ページには、NVMe ステータスが緑色で表示されます。

5 クラスタ内のすべてのネームスペースの詳細情報を表示するネームスペース管理ウィンドウを起動します。左側のメニューペインで、[Storage] > [NVMe] > [NVMe Namespaces] に移動します。以下のようにネームスペースを作成します。

5-1 [Create] をクリックします。

5-2 作成した SVM を選択します。

5-3 Advanced オプションを使用して、すべてのネームスペース名のプレフィックスとなる命名パターンを作成します。

5-4 [Naming Pattern] ダイアログボックスに、関連する詳細を入力します。

5-5 [Apply] をクリックします。

5-6 [Submit] をクリックしてネームスペースを作成します。

6 [Close] をクリックします。

図 A.7　ONTAP System Manager - NVMe ネームスペースの新規作成

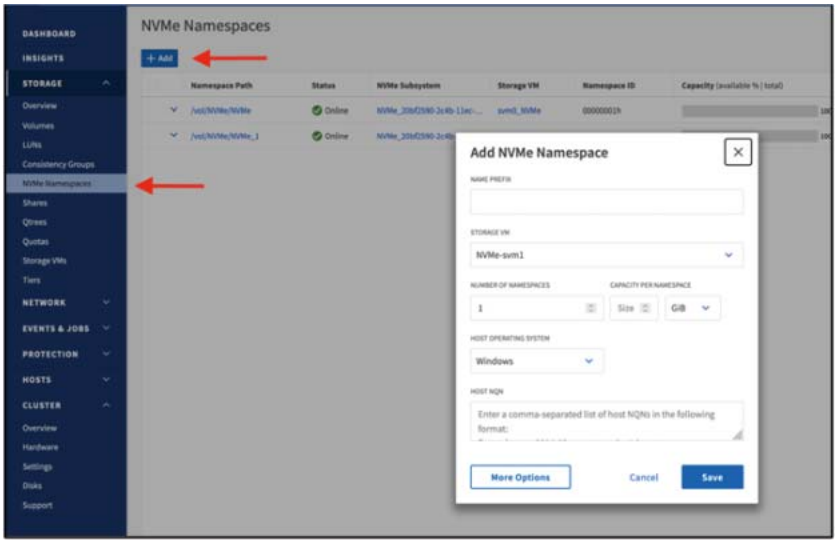


図 A.8　ONTAP System Manager - 新規 NVMe ネームスペースの表示

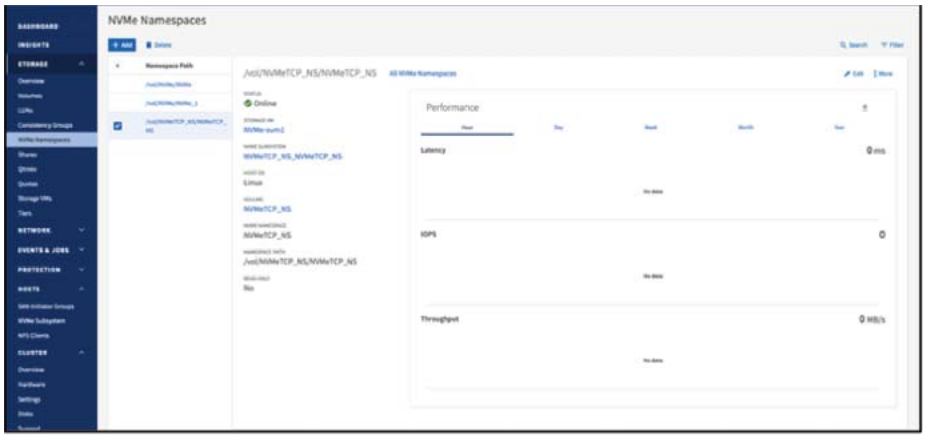
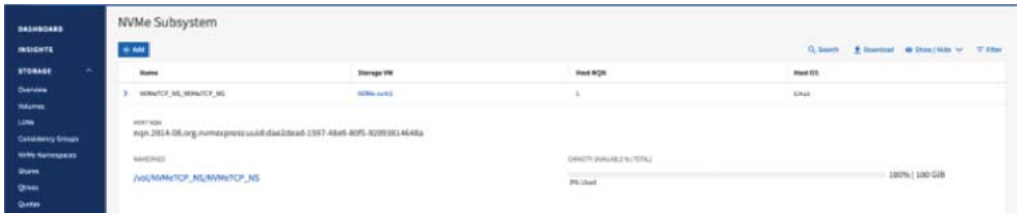


図 A.9　新規に作成された NVMe サブシステムの表示



付録 B
ONTAP NVMe/FC および NVMe/TCP CLI コマンド：初期設定と検出

■ ONTAP コントローラーの作業

手順 ▶▶▶

1 クラスタに NVMe/FC 対応のアダプタが取り付けられていることを確認します。

備 考

このアダプタは NVMe/FC には必要ですが、NVMe/TCP には不要です。

```
Cluster::> fcp adapter show -data-protocols-supported fc-nvme
 (network fcp adapter show)
           Connection Port   Admin  Operational
Node     Adapter Established Address  Status Status
------------ ------- ----------- --------- ------- ------------------------
AFF_1    1a   true    10100   up   online
AFF_1    1b   true    10200   up   online
2 entries were displayed.
```

2 NVMe/TCP LIF をホストする Ethernet ポートを表示します。

```
Cluster::> net port show
  (network port show)

Node: cluster-01
                                                      Speed(Mbps) Health
Port      IPspace      Broadcast Domain Link MTU  Admin/Oper  Status
--------- ------------ ---------------- ---- ---- ----------- --------
e0M       Default      Default          up   1500  auto/1000  healthy
e0a       Cluster      Cluster          up   9000  auto/100000
                                                               healthy
e1a       Cluster      Cluster          up   9000  auto/100000
                                                               healthy
e4a       Default      Default          down 1500  auto/-     -
e4b       Default      Default          down 1500  auto/-     -
e4c       Default      Default          down 1500  auto/-     -
e4d       Default      Default          down 1500  auto/-     -
<Output omitted>
e5b       Default      NVMeT            up   9000  auto/100000
                                                               healthy
19 entries were displayed.
```

3 NVMe トラフィックをホストする SVM を作成します。

```
Cluster::> vserver create -vserver nvme1
[Job 2831] Job succeeded:
Vserver creation completed.
```

4 作成された SVM を表示します。

```
Cluster::> vserver show
                Admin    Operational Root
Vserver   Type  Subtype  State   State    Volume   Aggregate
----------- ------- ---------- ---------- ----------- ---------- ----------
AFF     admin  -     -     -      -     -
AFF_1    node  -     -     -      -     -
AFF_2    node  -     -     -      -     -
nvme1    data  default  running  running   svm_root  AFF_2_SSD_1

4 entries were displayed.
```

5 SVM で許可されているプロトコルを表示します。

```
Cluster::> vserver show -vserver nvme1 -fields allowed-protocols

vserver allowed-protocols
------- -----------------
nvme1  -
```

6 NVMe プロトコルを追加します。

```
Cluster::> vserver add-protocols -vserver nvme1 -protocols nvme
```

7 SVM で許可されているプロトコルを表示します。

```
Cluster::> vserver show -vserver nvme1 -fields allowed-protocols
vserver allowed-protocols
------- -----------------
nvme1  nvme
```

8 NVMe サービスを作成します。

```
Cluster::> vserver nvme create -vserver nvme1
```

9 NVMe サービスのステータスを表示します。

```
Cluster::> vserver nvme show -vserver nvme1

     Vserver Name: nvme1
Administrative Status: up
```

10 NVMe/FC LIF を作成します。

```
Cluster::> network interface create -vserver nvme1 -lif fcnvme-node1-1a -role data -data-protocol
fcnvme -home-node node1 -home-port 1a
```

11 NVMe/TCP LIF を作成します。

```
network interface create -vserver nvme1 -lif lif_nvme1_182 -role data -data-protocol nvme-tcp -
home-node tme-a800-02 -home-port e5b -address 192.168.1.20 -netmask 255.255.255.0
```

12 新しく作成した LIF を表示します。

```
Cluster::> net interface show -vserver nvme1
 (network interface show)
      Logical  Status   Network      Current    Current Is
Vserver   Interface Admin/Oper Address/Mask    Node     Port  Home
----------- ---------- ---------- ------------------ ------------- ------- ----
nvme1
      fcnvme-node1-1a
             up/up  20:60:00:a0:98:b3:f7:a7
                           AFF_1     1a   true
      lif_nvme1_182
             up/up  192.168.1.20/24
                           AFF_1     e5b   true
```

13 LIF と同じノードにボリュームを作成します。

```
Cluster::> vol create -vserver nvme1 -volume nsvol1 -aggregate AFF_2_SSD_1 -size 50gb

Warning: You are creating a volume that, based on the hardware configuration, would normally have
the "auto" efficiency policy enabled. Because the effective cluster version is not 9.3.0 or
later, the volume will be created without the "auto" efficiency policy. After upgrading, the
"auto" efficiency policy can be enabled by using the "volume efficiency modify" command.
[Job 2832] Job succeeded: Successful   l
```

備 考

この警告は、`volume efficiency modify` コマンドを使用して作成するボリュームに「auto」性能を追加する必要があることを説明するものであり、無視しても問題ありません。

14 ネームスペースを作成します。

```
Cluster::> vserver nvme namespace create -vserver nvme1 -path /vol/nsvol1/ns1 -size 1GB -ostype linux
Created a namespace of size 1GB (1073741824).
```

15 サブシステムを作成します。

```
cluster1::> vserver nvme subsystem create -vserver nvme1 -subsystem mysubsystem -ostype linux
```

16 新しく作成されたサブシステムを表示します。

```
Cluster::> vserver nvme subsystem show -vserver nvme1
Vserver Subsystem  Target NQN
------- ------------ ------------------------------------------------------------
nvme1
    mysubsystem nqn.1992-08.com.netapp:sn.a6f7f76d40d511e8b3c900a098b3f7a7:subsystem.mysubsystem
```

■ ホストの作業

手順 ▶▶▶

1 ホストから NQN を取得します。

備 考

`hostnqn` 文字列の値は、nvme-cli パッケージのインストール時に `/etc/nvme/hostnqn` に自動的に設定されます。この値は永続します。この文字列はすでに一意に設定されています。したがって、Linux の `nvme gen-hostnqn` コマンドを使用して `hostnqn` 文字列を個別に生成する必要はありません。ホスト NQN が削除された場合は、Linux の `nvme get-host-nqn` ユーティリティを使用して生成できます。Linux ホスト NQN を永続化するには、`/etc/nvme/hostnqn` ファイルに追加します。

2 ホスト NQN を表示します。

```
SLES_host:~ # cat /etc/nvme/hostnqn
nqn.2014-08.org.nvmexpress:fc_lif:uuid:2cd61a74-17f9-4c22-b350-3020020c458d
```

■ ONTAP コントローラーの作業

手順 ▶▶▶

1 hostnqn 文字列をサブシステムに追加します。

```
Cluster::> vserver nvme subsystem host add -vserver nvme1 -subsystem mysubsystem -host-nqn nqn.1992-
08.com.netapp:sn.a6f7f76d40d511e8b3c900a098b3f7a7:subsystem.mysubsystem
```

2 ネームスペースをサブシステムにマッピングします。

```
Cluster::> vserver nvme subsystem map add -vserver nvme1 -subsystem mysubsystem -path /vol/nsvol1/ns1

Cluster::> vserver nvme namespace show -vserver nvme1 -instance

       Vserver Name: nvme1
      Namespace Path: /vol/nsvol1/ns1
           Size: 1GB
        Block Size: 4KB
         Size Used: 0B
          OS Type: linux
          Comment:
           State: online
       Is Read Only: false
       Creation Time: 4/15/2018 18:09:09
      Namespace UUID: 567fb229 -a05e-4a57-aec9-d093e03cdf44
   Restore Inaccessible: false
Node Hosting the Namespace: AFF_1
        Volume Name: nsvol1
        Qtree Name:
    Attached Subsystem: mysubsystem
       Namespace ID: 1
        Vserver ID: 89
```

◀◀◀

付録 C
ホスト構成情報

ホストの構成手順については、以下のサイトで確認できます。
https://www.fujitsu.com/jp/products/computing/storage/manual/

付録 D
LUN とネームスペース間の変換

■ LUN のネームスペースへの変換

手順 ▶▶▶

1 LUN を表示します。

```
tme-a700s-clus::> lun show
Vserver   Path                            State   Mapped   Type        Size
--------- ------------------------------- ------- -------- -------- --------
svm0      /vol/testLUN/testLUN                  online  mapped   linux        1GB
```

2 変換する LUN のマッピングを解除します。

```
tme-a700s-clus::> lun unmap -vserver svm0 -path /vol/testLUN/testLUN -igroup
new_15Mar21_tif5_igroup
```

3 LUN を変換します。

```
tme-a700s-clus::> vserver nvme namespace convert-from-lun -vserver svm0 -lun-path
/vol/testLUN/testLUN
```

4 ネームスペースを NVMe サブシステムにマッピングします。

```
vserver nvme subsystem map add -vserver svm0 -subsystem svm0_subsystem_909 -path
/vol/testLUN/testLUN
```

5 新しいネームスペースを表示します。

```
tme-a700s-clus::> vserver nvme namespace show
Vserver Path                              State      Size Subsystem       NSID
------- ------------------------------- ------- ------- ---------- ---------
Svm0    /vol/testLUN/testLUN              online      1GB svm0_subsystem_909 00000001h
```

◀◀◀

■ ネームスペースの LUN への変換

手順 ▶▶▶

1 新しいネームスペースを表示します。

```
tme-a700s-clus::> vserver nvme namespace show
Vserver Path                              State      Size Subsystem       NSID
------- ------------------------------- ------- ------- ---------- ---------
Svm0    /vol/testLUN/testLUN              online      1GB svm0_subsystem_909 00000001h
```

2 ネームスペースのマッピングを解除します。

```
vserver nvme subsystem map remove -vserver svm0 -subsystem svm0_subsystem_909 -path
/vol/testLUN/testLUN
```

3 ネームスペースを LUN に変換します。

```
lun convert-from-namespace -vserver svm0 -namespace-path /vol/testLUN/testLUN
```

4 LUN を igroup にマッピングします。

```
lun map -vserver svm0 -path /vol/testLUN/testLUN -igroup new_15Mar21_tif5_igroup -lun-id 20
```

5 LUN を表示します。

```
tme-a700s-clus::> lun show -vserver svm0
Vserver   Path                            State   Mapped   Type        Size
--------- ------------------------------- ------- -------- -------- --------
svm0      /vol/testLUN/testLUN            online  mapped   linux         1GB
```

付録 E
トラブルシューティング

NVMe/FC 障害のトラブルシューティングを行う前に、必ず富士通の仕様に準拠した構成になっていることを確認してください。確認が完了した後でホスト側の問題をデバッグするには、以下の手順を実行します。

NVMe/FC の lpfc 詳細ロギング

```
Here is a list of lpfc driver logging bitmasks available for NVMe/FC, as seen in
drivers/scsi/lpfc/lpfc_logmsg.h:
#define LOG_NVME 0x00100000 /* NVME general events. */
#define LOG_NVME_DISC 0x00200000 /* NVME Discovery/Connect events. */
#define LOG_NVME_ABTS 0x00400000 /* NVME ABTS events. */
#define LOG_NVME_IOERR 0x00800000 /* NVME IO Error events. */
```

`lpfc_log_verbose` ドライバ設定 (`/etc/modprobe.d/lpfc.conf` の `lpfc` 行に追加されます) を、`lpfc` ドライバの観点から NVMe/FC イベントをログに記録するための以前の値のいずれかに設定します。次に、`dracut-f` を実行して `initiramfs` を再作成し、ホストを再起動します。再起動後、以前の `LOG_NVME_DISC` ビットマスクを例にして以下の出力を確認し、詳細ロギングが適用されていることを確認します。

```
# cat /etc/modprobe.d/lpfc.conf
options lpfc lpfc_enable_fc4_type=3 lpfc_log_verbose=0x00200000
# cat /sys/module/lpfc/parameters/lpfc_log_verbose
2097152
```

一般的な問題については、以下の lpfc logging bitmask の値を推奨します。

- 一般的な NVMe 検出 / 接続イベント：0x00200000
- LIF/ ポート切り替えイベントなどのリンクバウンス時の FC-LS の検出問題に関連する lpfc ドライバイベント 0xf00083

一般的な nvme-cli エラーとその対処

このセクションでは、nvme-cli ユーティリティが `nvme discover`、`nvme connect`、および `nvme connect-all` の操作中に表示するエラーメッセージの一部について説明します。これらのエラーの考えられる原因とその対処について説明します。

エラーメッセージ

```
Failed to write to /dev/nvme-fabrics: Invalid argument.
```

- 考えられる原因
  このエラーメッセージは通常、構文が間違っている場合に表示されます。
- 対処
  以前の NVMe コマンドに正しい構文を使用してください。

● エラーメッセージ

```
Failed to write to /dev/nvme-fabrics: No such file or directory.
```

- 考えられる原因
  このエラーは、いくつかの問題が原因で発生します。一般的な原因には、以下のようなものがあります。
  - 誤った引数が以前の NVMe コマンドに渡された。
- 対処
  以前のコマンドに、正しい引数 (正しい WWNN 文字列、WWPN 文字列など) を使用していることを確認します。
  引数が正しいのにエラーが表示される場合は、/sys/class/scsi_host/host*/nvme_info の出力は正しく、NVMe イニシエータが Enabled と表示され、リモートポートセクションの下に NVMe/FC ターゲット LIF が正しく表示されます。
  例：

```
# cat /sys/class/scsi_host/host*/nvme_info
NVME Initiator Enabled
NVME LPORT lpfc0 WWPN x10000090fae0ec9d WWNN x20000090fae0ec9d DID x012000 ONLINE
NVME RPORT WWPN x200b00a098c80f09 WWNN x200a00a098c80f09 DID x010601 TARGET DISCSRVC ONLINE
NVME Statistics
LS: Xmt 0000000000000006 Cmpl 0000000000000006
FCP: Rd 0000000000000071 Wr 0000000000000005 IO 0000000000000031
Cmpl 00000000000000a6 Outstanding 0000000000000001
NVME Initiator Enabled
NVME LPORT lpfc1 WWPN x10000090fae0ec9e WWNN x20000090fae0ec9e DID x012400 ONLINE
NVME RPORT WWPN x200900a098c80f09 WWNN x200800a098c80f09 DID x010301 TARGET DISCSRVC ONLINE
NVME Statistics
LS: Xmt 0000000000000006 Cmpl 0000000000000006
FCP: Rd 0000000000000073 Wr 0000000000000005 IO 0000000000000031
Cmpl 00000000000000a8 Outstanding 0000000000000001
```

  - 対処
    ターゲット LIF が nvme_info 出力に上記のように表示されない場合は、/var/log/messages および dmesg 出力で疑わしい NVMe/FC 障害がないか確認し、それに応じて報告または修正します。

● エラーメッセージ

```
Failed to write to /dev/nvme-fabrics: Operation already in progress
```

- 考えられる原因
  このエラーメッセージは、コントローラーの関連付けまたは指定された操作がすでに作成されているか、作成中である場合に表示されます。このエラーは、インストールされている自動接続スクリプトの一部として発生する可能性があります。
- 対処
  なし。nvme discover の場合は、しばらくしてからこのコマンドを再実行してください。また、nvme connect および connect-all の場合は、nvme list を実行して、ホスト上でネームスペースデバイスがすでに作成され、表示されていることを確認します。

● エラーメッセージ

```
No discovery log entries to fetch
```

- 考えられる原因
  このエラーメッセージは通常、/etc/nvme/hostnqn 文字列が、ストレージアレイ上の対応するサブシステムに追加されていない場合に表示されます。このエラーは、誤った hostnqn 文字列がそれぞれのサブシステムに追加されている場合にも表示されます。

- 対処
  正しい /etc/nvme/hostnqn 文字列が、ストレージアレイ上の対応するサブシステムに追加されていることを確認します。vserver nvme subsystem host show コマンドを実行して確認します。

デバッグに必要なファイルとコマンド出力

問題が解決しない場合は、以下のファイルとコマンド出力を収集し、さらにトリアージを行うために富士通 SDK（保守契約が必要です）に送信してください。

```
 cat /sys/class/scsi_host/host*/nvme_info
/var/log/messages
dmesg
```

nvme discover 出力は、以下の通りです。

```
nvme discover --transport=fc --traddr=nn-0x200a00a098c80f09:pn-0x200b00a098c80f09 --host-
traddr=nn-0x20000090fae0ec9d:pn-0x10000090fae0ec9d
nvme list
```

付録 F
MCC IP での NVMe/FC の構成と設定

ONTAP 9.12.1 では、MCC IP 4 パッククラスタでの NVMe/FC の MCC IP サポートが追加されました。

MCC NVMe ホストのタイムアウト設定

ONTAP MCIP スイッチオーバー中に、プライマリクラスタサイトがセカンダリ / ピアクラスタサイトにスイッチオーバーしている間に、All Paths Down (APD) ウィンドウが表示されます。この APD ウィンドウが表示されている間、すべてのパスがダウンしているため、ホスト / クライアントはすべての NVMe ネームスペースデバイスにアクセスできなくなります。この APD ウィンドウが指定された間隔を超えると、Linux NVMe/FC ホストは処理を中断し、その上にあるアプリケーションに I/O エラーを返します。この動作は、リンク損失時の Linux NVMe/FC ホストの動作によって引き起こされ、`NVMe/FC dev_loss_tmo` と呼ばれるトランスポート層のパラメータによって制御されます。

NVMe プラットフォームのサポートと構成の制限

NVMe-oF プロトコルのサポートは、使用している ONTAP のバージョンに基づいて、プラットフォームおよび設定によって異なります。

NVMe FC LIF の制限

両方のクラスタの FC ポートを同じファブリックに接続し、ソフトゾーニングを使用する必要があります。これにより、接続されているポートに LIF が配置され、ホストはスイッチオーバー後に LIF にログインできるようになります。

詳細については、富士通マニュアルサイトに掲載の「ETERNUS AX/HX series MetroCluster ソリューションのアーキテクチャと設計」を参照してください。

付録 G
NVMe/TCP によるセキュアな認証の設定

手順 ▶▶▶

1 DH-HMAC-CHAP 認証を NVMe サブシステムに追加します。

```
vserver nvme subsystem host add- vserver svm_name
…
-subsystem subsystem
-host-nqn host_nqn
-dhchap-host-secret authentication_host_secret
-dhchap-controller-secret authentication_controller_secret
-dhchap-hash-function {sha-256|sha-512}
-dhchap-group {none|2048-bit|3072-bit|4096-bit|6144-bit|8192-bit}
```

2 DH-HMAC CHAP 認証プロトコルがホストに追加されていることを確認します。

```
vserver nvme subsystem host show
…
 [ -dhchap-hash-function {sha-256|sha-512} ] Authentication Hash Function
 [ -dhchap-group {none|2048-bit|3072-bit|4096-bit|6144-bit|8192-bit} ]
                Authentication Diffie-Hellman
                Group
 [ -dhchap-mode {none|unidirectional|bidirectional} ]
                Authentication Mode
```

3 DH-HMAC CHAP 認証プロトコルがコントローラーに追加されていることを確認します。

```
vserver nvme subsystem controller show
…
 [ -dhchap-hash-function {sha-256|sha-512} ] Authentication Hash Function
 [ -dhchap-group {none|2048-bit|3072-bit|4096-bit|6144-bit|8192-bit} ]
                Authentication Diffie-Hellman
                Group
 [ -dhchap-mode {none|unidirectional|bidirectional} ]
                Authentication Mode
```

◀◀◀

Fujitsu Storage
ETERNUS AX series オールフラッシュアレイ ,
ETERNUS HX series ハイブリッドアレイ
NVMe-oF を使用した最新の SAN の実装と構成

C140-0044-01Z3

発行年月　2023 年 6 月
発行責任　富士通株式会社

FUJITSU